Conception innovante et développement d'outils de conception d'ASIC pour Technologie Hybride CMOS / Magnétique

Grégory Di Pendina Di Pendina

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **NANO ELECTRONIQUE ET NANO TECHNOLOGIES**

Arrêté ministériel : 7 août 2006

Présentée par

« Grégory DI PENDINA »

Thèse dirigée par **« Kholdoun TORKI »** et
Co-encadrée par **« Guillaume PRENAT »**

préparée conjointement au sein
du **Laboratoire SPINTEC – CEA/CNRS/UJF** et du service **CMP**
dans **l'École Doctorale EEATS : Electronique, Electrotechnique, Automatique et Traitement du Signal**

Conception innovante et Développement d'outils de conception d'ASIC pour Technologie Hybride CMOS / Magnétique

Thèse soutenue publiquement le **« 19 Octobre 2012 »**,
devant le jury composé de :

Mr. Christian SCHAEFFER,
Professeur à Grenoble INP, Président du Jury
Mr. Pascal NOUET
Professeur à l'université de Montpellier II, Rapporteur
Mr. Amara AMARA
Professeur à l'Institut Supérieur d'Electronique de Paris, Rapporteur
Mr. Jean-Baptiste RIGAUD
Maître de conférences à l'ENSMSE, Examinateur
Mr. Kholdoun TORKI
Ingénieur de Recherche CNRS-CMP, Directeur de thèse
Mr. Guillaume PRENAT
Ingénieur - Chercheur CEA à SPINTEC, Encadrant de thèse
Mr. Laurent FESQUET
Maître de conférences à TIMA, Invité

Table des matières

Table des figures

Liste des tableaux

Résumé

Titre: Conception innovante et Développement d'outils de conception d'ASIC pour Technologie Hybride CMOS / Magnétique

Mots Clés: JTM: Jonction Tunnel Magnétique, MRAM: Magnetic Random Access Memory, technologies non volatile, circuit intégré full custom et/ou numérique, ASIC: Application Specific Integrated Circuit, kit de conception, fiabilité.

Depuis plusieurs années de nombreuses technologies non volatiles sont apparues et ont pris place principalement dans le monde de la mémoire, tendant à remplacer tout type de mémoire. Leurs atouts laissent à penser que certaines d'entre elles, et en particulier les technologies MRAM, pourraient améliorer les performances des circuits intégrés en utilisant leurs composants magnétiques, si connus notamment sous le nom de jonctions tunnel magnétiques, dans la logique. Pour évaluer ces éventuels gains, il faut être capable de concevoir de tels circuits. C'est pourquoi nous proposons dans ces travaux d'une part un kit de conception complet pour les flots de conception full custom et numérique, permettant de couvrir l'ensemble des étapes de conception pour chacun d'entre eux. Une partie de ce kit a servi à plusieurs partenaires de projets de recherche ANR, pour concevoir des démonstrateurs. Nous proposons également dans ce kit de conception un latch magnétique non volatil innovant ultra compact, pour lequel deux brevets d'invention ont été déposés, intégré à une flip-flop. Enfin, nous présentons l'intégration de composants magnétiques à deux applications, sécurité et faible consommation, ainsi qu'une étude qui montre que les gains en consommation statique peuvent être considérables.

Abstract

Title: ASIC Innovative design and Process Design Kit development for Hybrid CMOS / Magnetic Technology

Keywords: MTJ: Magnetic Tunnel Junction, MRAM: Magnetic Random Access Memory, non volatile technology, full custom and/or digital integrated circuit, ASIC: Application Specific Integrated Circuit, process design kit, low power, reliability.

For several years many non-volatile technologies have been appearing and taking place mainly in the memory world, aiming at replacing all kind of memory. Their assets let thinking that some of them, specially the MRAM technologies, could improve the integrated circuit performances, using their so called magnetic components in the logic, in particular the magnetic tunnel junctions. To evaluate the potential benefits, it is necessary to be able to design such a circuit. That is the reason why we are proposing a full design kit for both full custom and digital designs, allowing all the design steps. Part of this kit has been used by partners in research project to design demonstrators. We also propose in this kit an innovative ultra-compact magnetic latch, for which 2 patents have been deposited, integrated in a flip-flop. Finally, we present the integration of magnetic components for two applications, security and low power, as well as a case study which shows that the static consumption reduction can be huge.

Remerciements

- **Guillaume Prenat** - Evidemment, je dois le remercier tant il a été un soutien important, mais également pour tous les cours privés auxquels j'ai eu droit sur la spintronique, pour tous ses conseils, pour toute son aide et son support tout au long de ces 3 années.
- **Kholdoun Torki** - Pour m'avoir proposé de faire cette thèse et sans qui je n'aurais surement jamais eu l'opportunité de travailler dans le domaine de la spintronique.
- **Messieurs les membres du jury** - Mr. Christian Schaeffer, Mr. Pascal Nouet, Mr. Amara Amara, et Mr. Jean-Baptiste Rigaud, pour avoir accepté de présider, rapporter ou examiner ces travaux.
- **Olivier Goncalvès** - Doctorant dans l'équipe Design de Spintec, pour toute son aide à la programmation du testeur et pour le temps qu'il m'a consacré lors de la phase de test des démonstrateurs.
- **Jeremy Herault et Lucien Lombard** - Ingénieur fiabilité chez Crocus Technology, pour le support qu'ils nous ont fait sur le run Crocus, tant au niveau matériel en faisant l'interface pour nous attribuer un wafer, qu'au niveau technique sur les spécifications technologiques des lots auquels nous avons participé.
- **Bernard Courtois** - Pour m'avoir autorisé à mener en parallèle mes missions au sein du service CMP et mon travail de thèse.
- **Laurent Fesquet** - Pour nous avoir gentiment fourni les codes sources du processeur qu'il utilise en enseignement.
- **Sophie Dumont et Alexandre Chagoya** - Pour leurs conseils et astuces sur les outils de conception.
- **Azedine Manaa** - Pour avoir su mettre à profit ses connaissances en design de PCB et ainsi me permettre de fabriquer une carte de test pour les démonstrateurs.
- **Julien du T103** - Pour son aide "régulière" en informatique et pour sa bonne humeur chaque jour.
- **Ma femme et mon fils** - Pour leur patience et indulgence envers mes soirées peu souvent libres.
- **Mes parents et ma famille proche** - Pour leur soutien permanent et leurs encouragements pendant ces trois années (et celles d'avant...).

– **Les collègues** - Je ne citerai que Le Costaud, Le Grand et mon Chevreuil pour leur soutien moral et leur grande capacité à me faire décompresser les week-ends.

"La lumière voyage plus vite que le son. C'est bien pour ça que certains paraissent brillants jusqu'à ce que vous les entendiez parler"

"Les conneries c'est comme les impôts, on finit toujours par les payer - Michel Audiard, dialoguiste, scénariste, réalisateur français de cinéma, écrivain et chroniqueur de presse"

Introduction générale

Contexte

La spintronique et la microélectronique sont deux domaines relativement anciens à nos yeux. En effet, les premières découvertes sur le trajet d'électrons à travers des couches ferromagnétiques datent de 1936, d'abord suggérées par Mott [92] puis ensuite démontrées expérimentalement et décrites théoriquement à la fin des années 60 [36] [79]. Ce domaine a suscité beaucoup d'intérêt déjà à cette époque, ce qui a amené la découverte de la magnétorésistance géante en 1988, conjointement par les équipes d'Albert Fert, physicien français spécialiste de physique de la matière condensée, professeur émérite à l'Université Paris-Sud 11 en 2010 et directeur scientifique au sein de l'unité mixte de recherche CNRS/Thales, et par l'équipe de Peter Grünberg, physicien allemand. Cette découverte a donné naissance à un dispositif présentant beaucoup d'intérêts industriels et commerciaux. Il s'agit de la spin valve [32], configuration utilisée aujourd'hui dans les têtes de lecture et disques durs. La spintronique comme nous la connaissons aujourd'hui est basée sur le transport des électrons par effet tunnel magnétique démontré en 1975 par Jullière [78], suite aux travaux sur le transport des électrons par effet tunnel à travers un isolant, phénomène démontré en 1960 par Giaever [41], récompensé par le Prix Nobel de la Physique en 1973 [4]. Les découvertes de Jullière ont permis de définir un nouveau composant, la Jonction Tunnel Magnétique (JTM). Au cours de ces années, les caractéristiques de la spintronique ont intéressé le monde de la microélectronique, tout aussi convoité par la recherche scientifique. Dès 1947, date à laquelle sont apparues les premières technologies microélectroniques et le premier transistor [128], les premières mémoires ont vu le jour. Les progrès des technologies de circuits intégrés, tant au niveau des procédés de fabrication, qu'équipement, qu'outils de conception ou que de techniques de conception, ont été spectaculaires au cours des dernières décennies, suivant l'incontournable loi de Moore dictée en 1965 par Gordon E. Moore. Il constata alors que le nombre de transistors doublait chaque année dans les circuits intégrés [91],

puis il prédit en 1975 que le nombre de transistors des microprocesseurs doublerait tous les 2 ans. Au fil du temps, le chemin de ces deux domaines se sont croisés, par intérêts mutuels, ce qui a fait émerger une technologie hybride très répandues à ce jour, communément appelée MRAM, Magnetic Random Access Memory. En effet, c'est dans le domaine des mémoires que la spintronique a pris sa place au début de la rencontre de ces deux domaines. C'est d'ailleurs toujours le cas aujourd'hui, car la plupart des innovations dans ces technologies hybrides portent sur le sujet des mémoires. Les atouts principaux des jonctions tunnel magnétiques, à savoir la non volatilité, leur faible temps d'écriture et de lecture, leur endurance en comparaison aux autres technologies non volatiles, ainsi que son immunité aux radiations, font que leur intégration dans les mémoires augmentent les performances. Ces performances augmentent par ailleurs grâce aux progrès des procédés de fabrication magnétique, depuis les premières méthodes d'écriture FIMS (Field Induced Magnetic Switching) aux plus prometteuses et ambitieuses STT (Spin Transfert Torque), en passant par la méthode TAS (Thermally Assisted Switching).

En dehors des besoins propres aux mémoires, d'autres besoins se font ressentir de plus en plus et depuis plusieurs années dans le domaine de la microélectronique. En effet, notre vie quotidienne étant rythmée par une multitude d'appareils sans fil, téléphone, ordinateur, PDA, tablette et bien d'autres, les contraintes sur l'autonomie de ces appareils sont extrêmement fortes. Plus le procédé de fabrication est avancé en termes de finesse de gravure, plus la proportion de la consommation statique par rapport à la consommation dynamique devient importante. Les prédictions montrent que les proportions sont aujourd'hui de 65% de consommation statique et 35% de consommation dynamique, et que cette tendance devrait s'inverser, notamment grâce à des techniques et technologies nouvelles. La non volatilité de ces technologies hybrides CMOS / magnétiques laissent à penser qu'elles pourraient aider dans ce sens. Certaines technologies non volatiles, comme la mémoire Flash par exemple, ont déjà permis d'améliorer certains aspects. Cependant, avec l'avancée vers des procédés très submicroniques, elles présentent des inconvénients que n'ont pas les MRAM: la consommation, l'endurance et la miniaturisation pour n'en citer que quelques-uns. Pour franchir une étape supplémentaire sur les progrès de la microélectronique, il semble intéressant d'intégrer des composants magnétiques non volatils également dans la partie logique d'un circuit intégré, et non pas seulement au niveau de la mémoire d'un système mais également au niveau de la circuiterie. Nous avons souhaité à travers cette thèse répondre à ces interrogations afin d'évaluer les éventuels intérêts des technologies magnétiques émergentes et de voir si elles ont les atouts nécessaires pour répondre à ce type de besoin. Parallèlement à ces aspects de consommation, un

tel procédé de fabrication non volatil peut répondre à des besoins autres, tel que la sécurité dans les systèmes. En effet sauvegarder l'état d'un circuit ou d'une partie d'un circuit peut rendre certaines applications critiques plus sures. Enfin, n'oublions pas de mentionner que les jonctions tunnel magnétiques sont intrinsèquement immunes aux radiations et que leur intégration dans les circuits intégrés peut être un atout dans des applications spatiales.

Objectifs

A l'heure où ont débuté ces travaux de recherche, plusieurs thèses avaient déjà été menées sur le sujet des technologies MRAM: celle de Virgile Javerliac soutenue en 2006 intitulée *"Développement d'un modèle compact de la jonction tunnel magnétique de première génération et son intégration dans la réalisation d'architectures logiques reprogrammables hybrides magnétique-cmos"*, celle de Nicolas Bruchon soutenue en 2007 intitulée *"Evaluation, validation and design of hybrid CMOS"*, celle de Weisheng Zhao soutenue en 2008 intitulée *"Circuits logiques non-volatiles programmables utilisant des composants magnétiques"*, celle de Wei Guo soutenue en 2010 intitulée *"Compact Modeling of Magnetic Tunnel Junctions and Design of Hybrid CMOS/Magnetic Integrated Circuits"* et celle de Yoann Guillemenet soutenue en 2011 intitulée *"Logique magnétique pour l'exploration d'architectures reconfigurables"*. Toutes portent soit sur la modélisation des composants magnétiques, selon les différentes technologies MRAM qui ont vu le jour au fil des années, soit sur l'application à des circuits reprogrammable de type FPGA. Dans certaines d'entre elles, des démonstrateurs ont été conçus, plus ou moins complexes, mais aucune n'a proposé de circuit complet et complexe numérique, conçu selon des standards, aussi bien en termes de flot de conception que d'outils de conception. Cela a donc été un de nos objectifs, être capable de concevoir un circuit intégré de A à Z en utilisant les outils de conception industriels selon un flot tout à fait standard. Pour cela il était nécessaire d'une part de disposer et donc de concevoir des cellules standards innovantes spécifiques à cette technologie hybride CMOS / magnétique et de développer un kit de conception complet pour celle-ci. Ce PDK - Process Design Kit - a pour objectif de couvrir l'ensemble des étapes de conception, que ce soit pour un circuit suivant un flot full custom ou numérique, en intégrant des composants magnétiques au circuit. Enfin, un des objectifs était de faire une évaluation et une projection vers l'avenir des performances d'un circuit hybride complexe, notamment en termes de consommation, et de les comparer à un circuit CMOS conventionnel.

Plan du manuscrit

Dans le premier chapitre de ce manuscrit, une présentation des mémoires volatiles et non volatiles d'anciennes générations est faite. Suit un état de l'art des principales technologies non volatiles émergentes, et tout particulièrement celle des mémoires MRAM qui sont celles sur lesquelles nos travaux ont été menés. Ce chapitre se termine avec une synthèse sur les avantages et inconvénients de chacune d'entre elles. Le deuxième chapitre est consacré à la description des flots de conception de circuits intégrés, full custom et numérique, afin de bien appréhender la suite des travaux décrits. Le troisième chapitre décrit d'une part les quelques cellules proposées dans la littérature durant ces dernières années et d'autre part le latch magnétique innovant que nous proposons, pour lequel deux brevets d'invention ont été déposés, ainsi qu'une cellule de type flip-flop basée sur ce latch non volatil ultra compact. Le quatrième chapitre décrit l'ensemble des développements qui ont été faits pour la mise en place d'un kit de conception pour technologie hybride CMOS / Magnétique, ainsi que les flots de conception full custom et numériques spécifiques à celle-ci. Le cinquième chapitre est composé de deux parties. La première montre l'intérêt d'un procédé hybride pour des applications haute sécurité en termes de sauvegarde permanente de l'état d'un circuit intégré dans des composants non volatils. La seconde est une étude de consommation comparant un circuit numérique dans une application donnée pour des procédés CMOS standards et des procédés CMOS/Magnétique, selon plusieurs noeuds technologiques. Enfin, le sixième et dernier chapitre décrit les démonstrateurs que nous avons conçus et fabriqués pendant cette thèse, ainsi que les résultats de test.

Chapitre 1

Etat de l'art

1.1 Introduction

Le besoin de mémorisation de l'information s'est fait ressentir dans les applications électroniques il y a bien longtemps. Déjà au moment de l'apparition des premières technologies de microélectronique et du premier transistor en 1947 [128], les premières mémoires sont apparues. A chaque nouvelle génération ou nouvelle technologie, les avantages ont progressé, avec des besoins qui ont évolué aussi vite que la microélectronique elle-même. Celle-ci ayant suivi l'incontournable loi de Moore, la progression sur les mémoires s'est elle accélérée plus tard, au cours des 20 dernières années. En effet, en 1965, Gordon E. Moore constatait que la complexité des semiconducteurs proposés en entrée de gamme doublait tous les ans à coût constant depuis 1959 [91]. Il réévalua cette prédiction en 1975 disant que le nombre de transistors des microprocesseurs doublerait tous les 2 ans [39]. Par la suite plusieurs lois dérivées ont été massivement énoncées dans la littérature, parlant pour la plupart de 18 mois, relatif à la puissance, à la capacité ou à la vitesse, entre autres. Toujours est-il que l'évolution des mémoires a été spectaculaire également, bien qu'elle n'ait commencé qu'une vingtaine d'années plus tard. La figure 1.1 donne un état de l'art depuis les premières mémoires jusqu'à aujourd'hui, incluant non seulement les mémoires produites et fabriquées à grande échelle industrielle mais également les mémoires émergentes pour lesquelles beaucoup d'efforts sont faits au niveau de la recherche mondiale. Certaines sont d'ailleurs déjà commercialisées depuis quelques années.

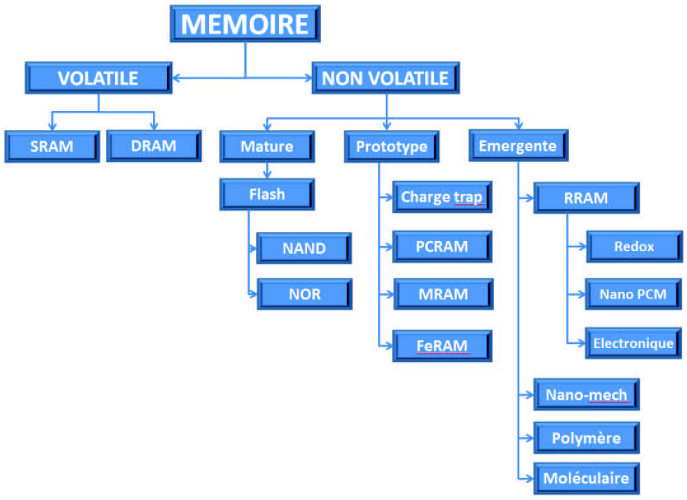


FIG. 1.1 – *Classification des mémoires*

1.2 Mémoires à semi-conducteurs

Les applications actuelles, de plus en plus sophistiquées et complexes, nécessitent de mémoriser des informations avec de fortes contraintes, soit en termes de quantité de données, soit en termes de rapidité, aussi bien en lecture qu'en écriture. Cette mémorisation peut être faite soit au niveau du circuit intégré lui-même, soit au niveau du système. On parle alors de hiérarchie mémoire d'un système. Cette hiérarchie est composée de mémoires volatiles, c'est à dire qui perdent leurs informations lorsque l'on coupe leur alimentation, comme par exemple les mémoires SRAM et DRAM d'un ordinateur de bureau. Cette même hiérarchie mémoire intègre également des mémoires non-volatiles, qui au contraire ne perdent pas les données lorsqu'elles ne sont plus alimentées. C'est le cas des clés USB par exemple ou des disques durs. La figure 1.2 illustrant cette hiérarchie montre également que plus les mémoires sont denses moins elles sont rapides, et inversement, plus elles sont rapides moins elles sont denses. La densité des mémoires s'exprime en F^2, soit le carré de la largeur de grille minimum de la technologie. Cette unité permet de comparer des densités sans tenir compte du noeud technologique.

Dans un système tel qu'un ordinateur, les mémoires très denses sont des disques durs (HDD: Hard Drive Disk) ou des mémoires Flash (SSD: Solid State Device). La mémoire dite centrale est une mémoire DRAM. Les caches de niveau 1 et niveau 2 sont des SRAMs. Les bancs de registres sont composés principalement de bascules de mémorisation, appelées flip-flops composées de 2 latchs SRAM. Enfin la partie d'un ordinateur qui effectue le calcul, le CPU (Central Processing Unit) utilise également

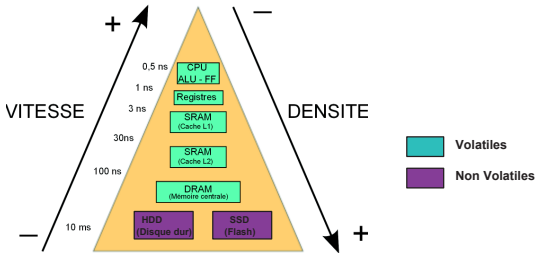


FIG. 1.2 – *Hiérarchie mémoire*

des éléments de mémorisation au niveau local, aussi à partir de flip-flops.

Parmi toutes les mémoires présentées sur la figure 1.1, nous allons dans ce premier chapitre détailler certaines d'entre elles, les plus courantes et les plus avancées à ce jour, en commençant par les mémoires volatiles, puis les non volatiles, pour terminer avec les mémoires non volatiles émergentes. L'état de l'art de la mémoire magnétique MRAM sera particulièrement développé car il s'agit du type de mémoire que nous avons utilisé pendant les travaux de cette thèse.

1.2.1 Cellule DRAM

La mémoire DRAM, pour Dynamic Random Acces Memory, a été inventée en 1966 par Dr. Robert Dennard au centre de recherche IBM Thomas J. Watson [99]. Un brevet US sera ensuite déposé en 1968 [31]. En 1970, la toute jeune entreprise Intel sort la première mémoire DRAM fabriquée, d'une capacité de 1K bit. C'est le modèle 1103. Cette mémoire sera le composant mémoire à semi-conducteur le plus vendu en 1972. La mémoire DRAM a pour principal avantage d'être très dense (environ 8 F^2) car elle n'est composée que d'un seul transistor de taille minimale et d'une capacité qui contient l'information, comme illustré sur la figure 1.3 qui montre sa structure.

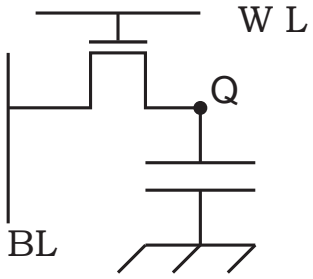


FIG. 1.3 – *Cellule mémoire DRAM*

Elle a comme principal inconvénient d'avoir un temps de rétention relativement court à cause des courants de fuite des capacités et doit alors être rafraichie très régu-

lièrement, toutes les 50 ms environ. Néanmoins, il a été démontré que la DRAM était capable d'avoir une durée de rétention bien plus importante à basse température [44], et sous certaines conditions jusqu'à plusieurs minutes [104]. Toutefois, l'information stockée est dégradée lors de la lecture, ce qui nécessite une réécriture systématique après chaque lecture. Cette mémoire est donc souvent utilisée dans les applications embarquées où la quantité de données et l'espace alloué sont des priorités. Elle est également utilisée comme mémoire centrale dans les ordinateurs mais aussi dans les consoles de jeux vidéo. C'est en revanche une mémoire qui a une consommation non négligeable en mode "standby" et qui est plus lente que la SRAM.

Pour la phase d'écriture, l'information à stocker, sous forme de signal logique '0' ou '1', est fournie par le signal bit line (BL). Lorsque la commande est activée à '1', gérée par le signal word line (WL), la capacité sera alors soit chargée si BL est à '1', soit déchargée si BL est à '0'. Il en est de même pour la phase de rafraichissement qui consiste au final à réécrire la même donnée que celle mémorisée. Pour la phase de lecture, le signal de commande WL est activé à '1'. Le niveau de tension aux bornes de la capacité est donc transmis sur le signal BL, lui-même connecté à une électronique spécifique de lecture, basé sur des amplificateurs de lecture, qui déterminent l'état logique du point mémoire.

1.2.2 Cellule SRAM

La mémoire SRAM, pour Static Random Access Memory, a été développée par Intel en 1969, dans le but de remplacer la mémoire traditionnelle des ordinateurs. Le premier produit fonctionnel à base de bipolaires rapides était un circuit de 64 bits, connu sous le nom de "modèle 3130 Shottky". Ensuite, en 1969 Intel a produit une SRAM à base de MOS d'une capacité de 256 K bits, qui sera le composant MOS semi-conducteur le plus vendu au monde en 1969, produit en grand volume [82].

Cette mémoire a pour principal avantage d'être très rapide et de consommer très peu pour maintenir l'information, tout en étant intrinsèquement stable. En effet, il suffit de quelques dizièmes de nanosecondes à quelques dizaines de nanosecondes pour l'écriture et pour la lecture. En revanche, elle a pour principal inconvénient d'être relativement peu dense (environ 90 F^2) car elle est composée de 6 transistors. Ces mémoires sont typiquement utilisées dans les niveaux de cache des systèmes à microprocesseur, où la vitesse est une priorité, ainsi que dans les applications nécessitant des accès mémoire rapides et fréquents. La figure 1.4 montre son architecture très largement répandue.

Cette structure est composée de 2 inverseurs montés tête-bêche, encadrés par 2 transistors d'accès. L'information à stocker est fournie par le signal bit line (BL) et

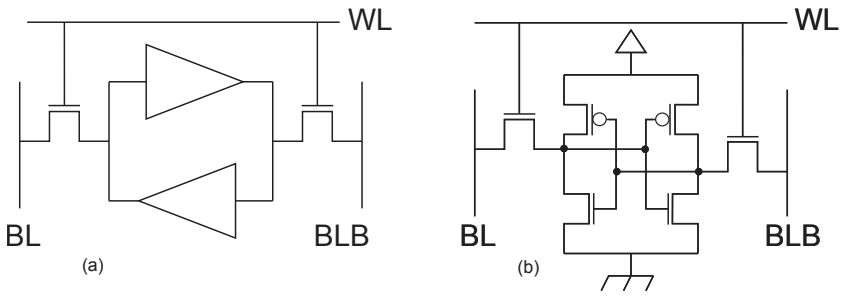


FIG. 1.4 – *Cellule mémoire SRAM*

son complément par le signal bit line bar (BLB). Il y a donc dans une cellule SRAM en permanence 1 seul bit stocké mais sur 2 noeuds de stockage complémentaires l'un de l'autre: l'information ('1' par exemple) et son complément ('0' par exemple). La commande est gérée simultanément pour les 2 transistors d'accès par le signal word line (WL). Lors de la lecture, les 2 signaux BL et BLB sont connectés à un amplificateur de lecture et sont préchargés à vdd (figure 1.5). Ainsi, lorsque les transistors d'accès sont activés par le signal WL, la ligne bit line connectée au bit stocké '0' va se décharger à travers le transistor d'accès dans un des 2 inverseurs. Ceci provoque une chute de tension qui sera détectée par l'amplificateur de lecture. Cette lecture se fait donc de façon différentielle entre les signaux BL et BLB.

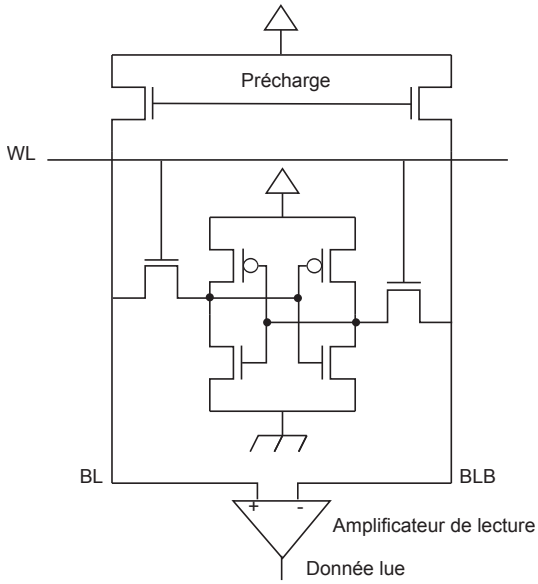


FIG. 1.5 – *Schéma de lecture d'une SRAM*

Au-delà des besoins de mémoires denses telles que les DRAM ou rapides telles que les SRAM, le besoin de mémoire non volatile est aussi très important dans beaucoup de systèmes. Parmi les mémoires non volatiles utilisées actuellement dans les systèmes de la microélectronique, on retrouve entre autres les ROM, les PROM, les EPROM, les EEPROM et les Flash. Nous allons présenter dans la suite de ce chapitre chacune d'entre elle, puis nous aborderons enfin les mémoires émergentes.

1.2.3 Cellule ROM

La mémoire ROM, pour Read Only Memory. La toute première ROM a vu le jour en même temps que les premières technologies à semi-conducteurs, au début du XX^{e} siècle. Elle a pour principal avantage d'être non volatile et pour principal inconvénient d'être à programmation unique. Le composant de base d'une telle mémoire est soit une diode, soit un transistor. La présence ou l'absence d'un composant dans une matrice permet de coder soir un '1' soit un '0 'logique, comme illustré sur la figure 1.6. En effet, la programmation se fait lors de la fabrication et est irréversible, les données écrites sont alors figées. Le procédé de fabrication est relativement compliqué et coûteux, dans la mesure où il nécessite des masques spécifiques. Le temps d'accès en lecture est d'environ 150 ns.

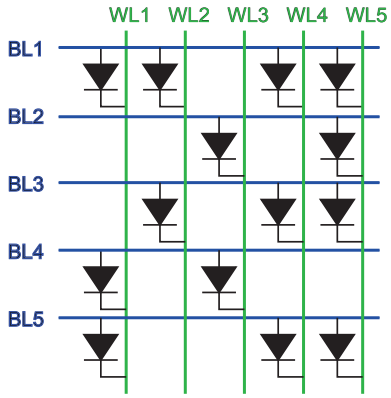


FIG. 1.6 – *Cellule mémoire ROM*

Les mémoires ROM sont généralement utilisées dans les ordinateurs et contiennent des instructions qui sont souvent utilisées par le processeur, lors du démarrage du système par exemple, ou dans des applications nécessitant d'exécuter un même programme régulièrement. Son contenu est dans ce cas un programme préalablement défini basé sur le jeu d'instruction du processeur. La ROM peut contenir par exemple le BIOS d'un système d'exploitation d'un PC, programme permettant de piloter les interfaces d'entrée-sortie principales du système.

1.2.4 Cellule PROM

La mémoire PROM, pour Programmable Read Only Memory. La PROM a été inventée en 1956 par Wen Tsing Chow, alors employé chez "Arma Division of the American Bosch Arma Corporation". Les toutes premières PROM fonctionnelles sont apparues dans les années 60 [101] puis ont été mises au point à la fin des années 70 par la firme Texas Instruments. Elle a pour principal avantage d'être programmable par l'utilisateur en plus de sa non volatilité. L'écriture des données dans la mémoire se fait électriquement, principalement par destruction de fusibles en appliquant une tension forte. Le schéma de principe est illustré en figure 1.7. Cette opération d'écriture est tout à fait irréversible. Les fusibles ou diodes ainsi grillés correspondent à des '0' logiques, les autres correspondent à des '1' logiques. Ce type de programmation reste un inconvénient majeur, car il s'agit d'une programmation unique et la PROM ne peut donc être écrite qu'une seule fois.

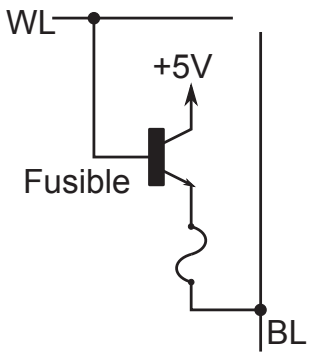


FIG. 1.7 – *Cellule mémoire PROM à fusibles*

Les PROM sont souvent utilisées dans des applications du type téléphonie mobile, consoles de jeu vidéo, composants pour le médical, ainsi que dans certaines applications RF de type RFID Tags (code à barres, passeports, carte de transport, carte de paiement)[53].

1.2.5 Cellule EPROM

La mémoire EPROM, pour Erasable Programmable ROM. En septembre 1970, l'EPROM a été annoncée en interne chez Intel. Elle a tout d'abord suscité beaucoup de scepticisme, pour ensuite amener beaucoup d'enthousiasme lorsque D. Frohman a reçu le titre de "Best Paper Award" à la conférence "International Solid-State Circuits Conference" (ISSCC) à Philadelphie en 1971 [101] [38]. Cette mémoire peut être réécrite plusieurs fois, ce qui est un avantage considérable par rapport au ROM

et PROM précédentes. Cependant l'effacement se fait par exposition du composant aux rayons ultraviolets, d'une longueur d'onde inférieure à 400 nm. Une exposition à la lumière solaire pendant 1 an ou à la lumière intérieure fluorescente pendant 3 ans peut aussi engendrer l'effacement de ces mémoires. Ce type d'effacement par UV nécessite d'une part un équipement spécifique, mais également une manipulation physique du composant par un opérateur. De plus le temps d'exposition recommandé est de 20 à 30 minutes pour des UV de 253.7 nm et une puissance de 15 W-s/cm^2, à une distance d'environ 2.5 cm [48]. L'effacement n'est donc pas dynamique dans son application. Pour rendre possible cette exposition aux UV, les circuits EPROM sont montés dans des boitiers à fenêtre transparente qui permet aux rayons ultraviolets d'atteindre le circuit intégré, comme illustré sur la figure 1.8.

FIG. 1.8 – *Cellule mémoire EPROM à fenêtre*

On retrouve les mémoires EPROM principalement dans les microcontrôleurs ou dans des imprimantes laser par exemple, surtout avant l'arrivée de nouvelles mémoires non volatiles telles que les EEPROM ou Flash.

1.2.6 Cellule EEPROM

La mémoire EEPROM, pour Electrically Erasable Programmable ROM. Les premiers développements des EEPROM ont débuté en 1978, par George Perlegos alors employé chez Intel. Il développa la première EEPROM "Intel 2816", puis créa la société "Seeq technology" en 1981, date à laquelle les premières EEPROM complètement fonctionnelles sont apparues [74]. Le fusible de la PROM est ici remplacé par un transistors M-NOS: Metal Nitride Oxide Semiconductor, comme le montre la figure 1.9. Les charges sont piégées à l'interface du Nitride, et ne sont pas perdues lorsque l'alimentation est coupée [81]. Ces mémoires sont facilement effaçables par impulsion électrique et l'on peut stocker de nouvelles données ainsi facilement de façon dynamique.

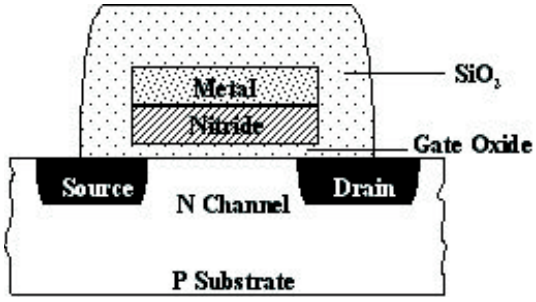


FIG. 1.9 – *Cellule mémoire EEPROM MNOS*

Les EEPROM sont ainsi fréquemment utilisées pour la mémorisation de données de configuration, dans certains ordinateurs ou dans les systèmes n'ayant pas besoin de stocker une grande quantité de données.

1.2.7 Cellule Flash

La mémoire Flash est la plus jeune technologie des mémoires non volatile non émergente avec moins de 30 ans de commercialisation [6]. Elle a été inventée par Dr. Fujio Masuoka alors employé chez Toshiba, dans les années 1980. Ces travaux ont été présentés par Dr. Masuoka en 1984 à la conférence "International Electron Devices Meeting (IEDM)" à San Francisco [83]. Il existe 2 types de mémoire Flash: les NOR et les NAND, toutes les deux ayant le même inventeur.

La flash NOR a été la première à être développée commercialement par Intel en 1988 [114]. Elle a pour caractéristique d'être bien plus lente en écriture et moins dense que son homologue. En effet les temps d'effacement et d'écriture sont longs, jusqu'à de l'ordre de la seconde selon sa longueur [30]. En revanche elle possède une interface d'adressage permettant un accès aléatoire et rapide à n'importe quelle position, ce qui fait qu'elle est plus rapide que la Flash NAND en lecture. La mémoire Flash NOR est adaptée à l'enregistrement de données informatiques destinées à être exécutées directement à partir de cette mémoire. Cette caractéristique est appelée XIP (eXecute In Place). La mémoire NOR est particulièrement bien adaptée pour contenir le système d'exploitation par exemple (OS: Operating System), dans les téléphones portables, principal marché des Flash NOR, mais aussi dans les décodeurs télés, les cartes mères ou leurs périphériques du type imprimantes, appareils photos, etc. Du fait de son coût, bien plus élevé que celui de la Flash NAND et de sa densité limitée, elle n'est en général pas utilisée pour le stockage de masse. L'endurance de ce type de composant est de l'ordre de 100 écritures / lectures dans le cas d'une intégration "on-chip" [1] mais plus typiquement 10^5 cycles, voire 10^6 [11]. Sa structure

est présentée sur la figure 1.10 montrant l'analogie avec une porte logique NOR.

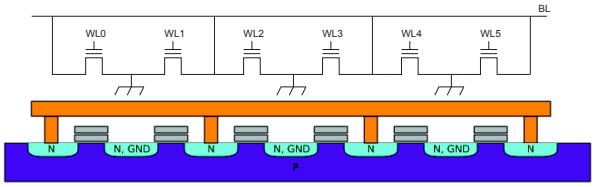


FIG. 1.10 – *Cellule mémoire Flash NOR*

La flash NAND a elle été annoncée par Toshiba en 1987 au meeting "International Electron Devices". Elle a ensuite été commercialisée par Toshiba en 1989. Elle a pour propriété d'être intrinsèquement plus rapide en lecture et en écriture que son homologue, de l'ordre de la milliseconde. Toutefois son interface d'entrée / sortie n'autorise que l'accès séquentiel. Cela tend à limiter, au niveau du système, sa vitesse effective de lecture, et à compliquer le démarrage direct à partir d'une mémoire NAND. De ce fait elle est moins bien adaptée que la NOR pour exécuter du code machine. La Flash NAND est en revanche plus dense grâce à la réduction du nombre de connexions vers la masse et vers les bit lines, ceci malgré le nombre de transistors plus important. Elle est donc adaptée au stockage de masse faible coût, comme les cartes mémoires ou les clés USB, ainsi que dans les applications d'électronique embarquée, appareils photos numériques, les téléphones mobiles, ordinateurs et baladeurs portables entre autres. Le nombre de cycle d'écriture est du même ordre de grandeur que la Flash NOR, soit environ 10^5. La raison d'avoir une endurance si faible est que les écritures nécessitent l'application de tensions plus élevées que la simple lecture, ce qui endommage peu à peu la zone écrite, bien qu'il existe des techniques de répartition de l'usure, par des procédés variant selon les constructeurs. Les lectures, même répétées, ne lui causent en revanche aucun dommage. Sa structure est présentée sur la figure 1.11 montrant l'analogie avec une porte NAND.

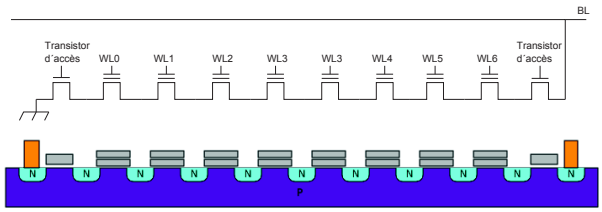


FIG. 1.11 – *Cellule mémoire Flash NAND*

Ces mémoires Flash remplacent aujourd'hui les EPROM et EEPROM et sont très répandues dans les systèmes actuels. Elles ont l'avantage d'être aussi denses que

les EPROM et d'avoir la souplesse de programmation des EEPROM. Dans cette technologie plus récente, les transistors M-NOS des EEPROM sont remplacés par des transistors double grilles, ce qui est cependant un inconvénient majeur du point de vue de la fabrication. Le procédé de fabrication est relativement complexe car comme on le voit sur la figure 1.12 cette cellule nécessite d'avoir 2 grilles l'une au-dessus de l'autre. Celle du haut est utilisée pour contrôler l'écriture, celle du bas est une grille flottante qui sert au codage de l'information. L'information est stockée grâce au piégeage d'électrons dans cette grille flottante, soit par l'injection d'électrons chauds soit par effet tunnel. L'écriture est relativement lente puisqu'il faut environ 50us pour stocker une donnée, nécessitant de plus une tension élevée, de l'ordre de 10V. Ces 2 aspects sont également des inconvénients majeurs.

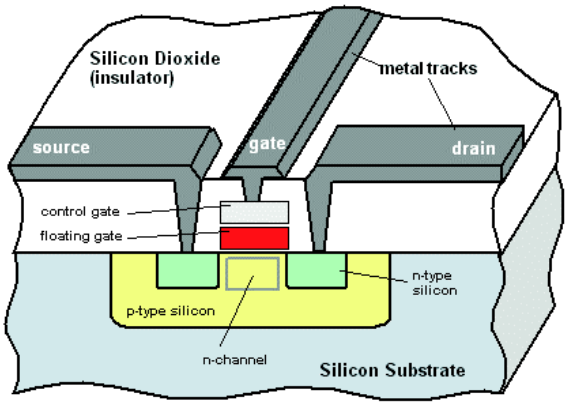


FIG. 1.12 – *Cellule mémoire Flash à double grille*

Afin d'augmenter la densité des Flash NAND une architecture multi niveaux (MLC) est utilisée [112]. Aujourd'hui se pose la question sur la capacité de cette mémoire à être miniaturisée en dessous des noeuds 25nm. En effet, les difficultés liées à la lithographie, aux matériaux et à la fabrication sont grandissants [87].

1.3 Mémoires émergentes non volatiles

Les mémoires à semi-conducteurs sont maintenant utilisées depuis plus de 40 ans [24]. Leur densité a quadruplé environ tous les 3 ans, de 1 K-bits en 1972 à plus de 1G-bits en 2012. Les besoins qui, historiquement, laissaient place à un compromis entre la vitesse et la consommation, demandent aujourd'hui à la fois des mémoires rapides et qui consomment très peu. Alors que le prix a chuté de moins de 1 dollar par tranche de 100 bits à moins de 1 dollar par tranche de 100 Mbits, le procédé de fabrication en production est devenu si complexe et coûteux que seules les grandes

sociétés de fabrication de semi-conducteurs sont capables de supporter ce marché. Les circuits ont atteint des géométries si petites que les théories de la physique classique ne suffisent plus à décrire leur comportement qui est dicté par des phénomènes quantiques. Les paramètres qui étaient négligeables jusque-là ne le sont plus. Bien que les technologies de mémoires à semi-conducteur aient encore plusieurs années de vie devant elles, le monde de la microélectronique est en train de s'intéresser très fortement aux technologies de mémoires émergentes, et ce depuis déjà quelques années.

La technologie idéale serait donc celle qui allie les avantages de toutes les mémoires les plus utilisées dans les systèmes actuels, à savoir la vitesse de la SRAM, la densité de la DRAM et la non volatilité de la Flash. De plus, elle devrait être miniaturisable pour pouvoir suivre l'évolution des technologies CMOS à semi-conducteurs. Cette mémoire permettrait d'éviter de nombreux transferts de données entre différents circuits, ce qui améliorerait d'une part les performances et la fiabilité, et réduirait les coûts d'autre part, financiers et énergétiques.

Parmi ces mémoires émergentes, on en dénombre une importante quantité comme nous l'avons vu sur la figure 1.1. Cependant nous n'en présenterons que quelques-unes dans ce manuscrit, les plus répandues à ce jour, à savoir les mémoires FeRAM, PCRAM, RedOx-RRAM et MRAM.

1.3.1 Mémoire FeRAM

La mémoire FeRam, pour Ferro-Electric RAM. Cette mémoire a été proposée par un étudiant de MIT, Dudley Allen Buck, pendant ses travaux de thèse "Ferroelectrics for Digital Information Storage and Switching", alors publiée en 1952 [23]. Les premiers développements ont commencé dans les années 80 et la FeRAM est apparue en 1983. Elle a fait concurrence aux EEPROM. Depuis 2001 ces mémoires sont en production dans leur version "2 transistors - 2 capacités" (2T2C), son architecture étant présentée en figure 1.13 [24]. L'information est stockée dans une capacité comme dans une DRAM, mais dans le cas de la FeRAM un film ferroélectrique est utilisé en guise de diélectrique, ce qui lui apporte la non volatilité. Sa vitesse d'écriture de l'ordre de 50ns et sa taille lui permettent d'être compétitive avec la SRAM, en plus de consommer moins. En revanche, elle reste moins dense que la DRAM. Elle est aussi bien plus rapide que la Flash en écriture mais avec des capacités de stockage plus petites. Son endurance est plus importante, 10^{16} pour des composants 3.3 V. Son procédé de fabrication est tel que le coût de fabrication reste encore assez élevé aujourd'hui.

Actuellement, une version "1 transistor - 1 capacité" (1T1C) a été développée,

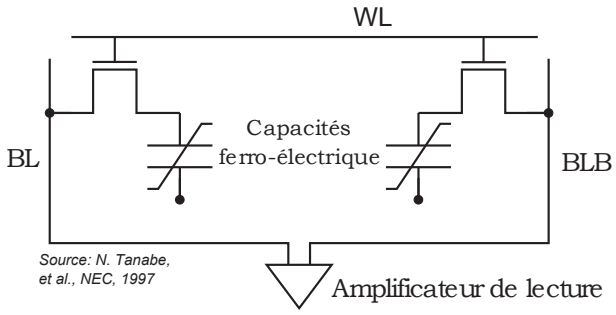


FIG. 1.13 – *Cellule mémoire FeRAM 2T2C*

comme illustré sur la figure 1.14[8]. Celle-ci est concurrente à la DRAM et à la Flash en termes de densité. Aujourd'hui, les produits de Fujitsu intègrent des composants LSI à base de FeRAM. Fujitsu, qui a été le premier à intégrer des FeRAM dans des micro-ordinateurs en 1988 [67], a fourni en 2011 plus de 1 milliard de composants incluant des FeRAM, leader dans le développement et la fabrication de ces mémoires.

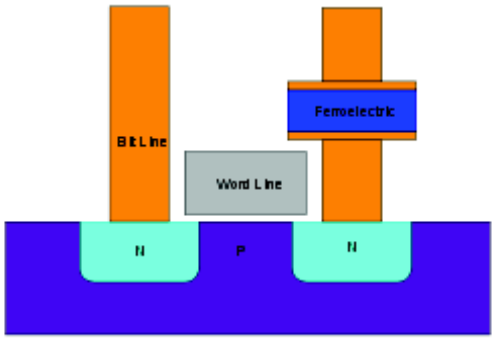


FIG. 1.14 – *Cellule mémoire FeRAM 1T1C*

1.3.2 Mémoire PCRAM

La mémoire PCRAM, pour Phase Change RAM. Dans les années 60, Stanford R. Ovshinsky de "Energy Conversion Devices" a été le premier à explorer les propriétés du verre à chalcogénures en vue de technologie mémoire potentielle. En 1969, Charles Sie publie une dissertation à l'université de l'état de l'IOWA aux Etats Unis [108] [106]. Il décrit et démontre la faisabilité d'une mémoire à changement de phase en intégrant des films de chalcogénures à une matrice de diodes. Une étude cinématographique montre en 1970 que le mécanisme de mémoire à changement de phase

dans le verre à chalcogénures implique la croissance de filament cristallin par effet de champ électrique induit [107]. En septembre 1970, Gordon Moore, co-fondateur d'Intel, publie un article sur cette technologie. Cependant, les problèmes de qualité des matériaux et de consommation ont empêché toute commercialisation de cette technologie. Bien plus tard, en août 2004, Nanochip met en place un accord de licence pour la technologie PCRAM pour l'utilisation dans les MEMS - Micro Electrical Mechanical Systems. En septembre 2006, Samsung annonce une mémoire prototype de 512 Mb (64Mo) utilisant des switchs à diode [2]. Enfin, en 2011, IBM annonce qu'ils ont créé une mémoire à multi changement de phase stable, fiable avec de très hautes performances [7].

La mémoire PCRAM est une mémoire à accès aléatoire à changement de phase d'un matériau particulier, pouvant se trouver sous deux formes distinctes, vitreuse ou cristalline. Le matériau utilisé est un verre à chalcogénures. Le passage de l'une à l'autre phase est assuré par une élévation en température par impulsion électrique, l'opération représentant l'écriture d'un '0' ou d'un '1'. Ce procédé est également employé pour les disques réinscriptibles, CD-RW ou DVD-RW. Un faible courant électrique assure la lecture qui dans ce cas se fait de façon optique à partir d'un rayon laser, mieux réfléchi par la forme cristalline que par la forme vitreuse. La résistivité étant beaucoup plus forte pour la forme vitreuse que pour la forme cristalline, le courant sortant permet de distinguer les deux états. La figure 1.15 [49] est une vue de coupe de 2 cellules PCRAM qui sont dans 2 états différents. L'une est dans un état à faible résistance cristalline et l'autre est dans un état de haute résistance amorphe. Le temps d'écriture est de l'ordre de 10 ns pour passer de l'état cristallin à l'état amorphe. En revanche il est d'environ 50 ns pour passer de l'état amorphe à l'état cristallin. Cette asymétrie importante dans l'écriture des 2 états peut être vue comme un inconvénient par rapport aux performances relatives aux aspects de timing.

Intel, Samsung et IBM ont présenté des modèles de PCRAM. Son domaine d'utilisation est celui de la mémoire Flash, avec l'avantage d'une vitesse d'écriture trente fois plus élevée et une durée de vie dix fois plus grande (chiffres annoncés par Samsung) [68]. Elle a également l'avantage d'être intrinsèquement immune aux radiations du fait que la technologie n'implique pas de transport de charge [127].

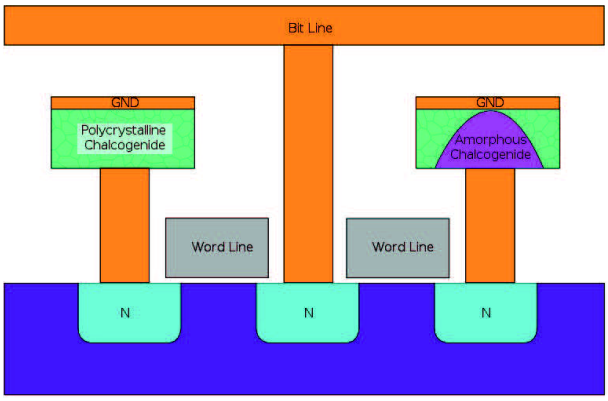


FIG. 1.15 – *Cellule mémoire PCRAM*

1.3.3 Mémoire RedOx-RRAM

La mémoire RedOx-RRAM pour Reduction Oxidation Resistive RAM fait partie de la famille des mémoires à commutation de résistances, appelée ReRAM ou encore Memristor, pour "memory resistor". L'effet memristor a été prédit et décrit en 1971 par Leon Chua de UC Berkeley [29]. Depuis 1971, la memristor était un composant hypothétique, aucun exemple physique n'étant connu. En avril 2008, soit 37 ans plus tard, une implémentation physique de la memristor a été reportée dans le journal "Nature" par une équipe de chercheurs des laboratoires HP conduite par R. Stanley Williams [119] [111] [77]. Une memristor est un composant passif qui stocke efficacement l'information car la valeur de sa résistance électrique change, de façon permanente, lorsqu'un courant est appliqué. Sa structure est un empilement de type MIM, Metal Isolant Metal. Là où une résistance classique apporte une valeur stable de résistance, une memristor peut avoir une valeur élevée de résistance lorsqu'elle est traversée par un courant dans un sens, interprétable comme un '1' logique, et une valeur faible lorsqu'elle est traversée par un courant dans le sens opposé, qui peut être interprétée comme un '0' logique [5]. Ainsi, une donnée peut être enregistrée et réécrite par un courant de contrôle dépendant du sens, comme illustré sur la figure 1.16. On parle alors de structure bipolaire. Dans le cas des structures unipolaire, le changement de résistance se fait par application soit d'une tension élevée, soit d'une tension faible.

La mémoire RedOx-RRAM est une bonne candidate pour être la mémoire volatile de demain [122]. En effet elle est plus rapide que les PCRAM, nécessite une tension plus faible que la flash donc plus intéressante du point de vue de la consommation. Elle est également potentiellement aussi rapide que la DRAM. Enfin, le procédé de fabrication étant relativement simple, elle semble être miniaturisable jusqu'à

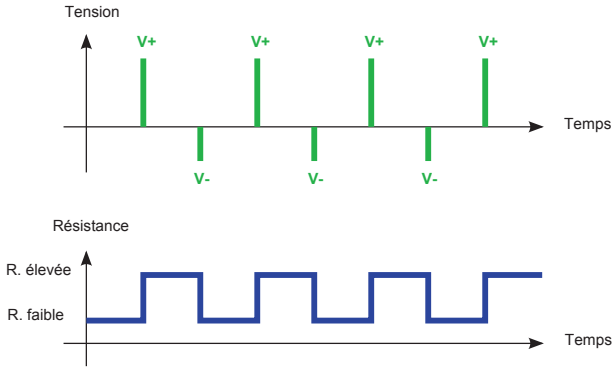


FIG. 1.16 – *ReRAM bipolaire*

des technologies 8nm. Cependant, son endurance de l'ordre de 10^8 est une faiblesse en comparaison avec d'autres technologies émergentes. La RedOx-RRAM n'est pas la seule piste suivie dans les laboratoires de recherche et d'autres technologies de mémoires émergentes ont également de très bons atouts pour devenir la mémoire universelle de demain.

1.3.4 Mémoire MRAM

Les mémoires MRAM, pour Magnetic RAM. Nos travaux étant basés sur cette technologie de mémoire, les paragraphes suivants permettent d'introduire plusieurs notions fondamentales, tel que la spintronique, la MagnétoRésistance Géante (GMR), la MagnétoRésistance Tunnel (TMR) ou encore les jonctions tunnel magnétiques (JTM). Chacune de ces définitions sont donc décrites et détaillées dans la suite de ce chapitre, dans lequel une large section est consacrée aux mémoires MRAM.

1.3.4.1 La spintronique

Qu'est que la spintronique ? Un électron est caractérisé par 3 éléments fondamentaux: sa charge électrique, sa masse et son spin. Le principe de la spintronique est d'utiliser non seulement la charge des électrons mais également leur spin, c'est à dire leur moment angulaire intrinsèque. Cette appellation fait l'analogie avec le moment cinétique d'une masse en rotation, bien que ce modèle soit inexact dans le cas du spin qui est une propriété purement quantique n'ayant pas d'équivalent en mécanique classique. A ce moment angulaire est associé un moment magnétique élémentaire. Dans le vide, on définit l'excitation magnétique $\vec{B_0}$ et le champ magnétique $\vec{H}$, reliés par la permitivité magnétique du vide μ_0 :

$$\vec{B_0} = \mu_0 \vec{H} \tag{1.1}$$

Hors vide, l'excitation magnétique d'un matériau dépend de la somme du champ magnétique au sein du matériau et de l'aimantation, suivant la formule suivante:

$$\vec{B} = \mu_0 \left(\vec{H} + \vec{M} \right) \tag{1.2}$$

Dans le cas de matériaux dits ferromagnétiques, les interactions entre les spins voisins tendent à les aligner dans la même direction et selon la même orientation, ce qui crée une aimantation spontanée macroscopique au sein du matériau. Son aimantation $\vec{M}$ est alors non nulle car elle est la somme de tous les moments magnétiques. Dans le cas des matériaux antiferromagnétiques, l'interaction entre les spins voisins tendent à les aligner dans la même direction, mais avec des orientations opposées, résultant en une aimantation globale nulle. Enfin, les matériaux conducteurs classiques tels que l'aluminium ou le cuivre que l'on retrouve dans les circuits microélectronique sont dits paramagnétiques. En magnétisme cela désigne le comportement d'un milieu qui ne possède pas d'aimantation spontanée. Son aimantation $\vec{M}$ est donc nulle car tous les moments magnétiques sont désordonnées. La figure 1.17 illustre l'organisation des moments magnétiques pour un matériau paramagnétique (a), celle pour un matériau ferromagnétique (b) et celle pour un matériau anti-ferromagnétique (c).

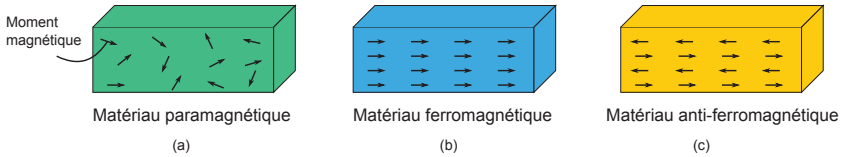


FIG. 1.17 – *Orientation des moments magnétiques*

Le spin de l'électron est dit soit up, si le l'électron "tourne" dans le sens horaire, soit "down" si l'électron "tourne" dans le sens antihoraire. La figure figure 1.18 montre un exemple de configuration où les électrons ayant leur spin "up" sont largement majoritaires par rapport à leurs homologues. Il s'agit d'une propriété dite d'aimantation spontanée [56].

Lorsqu'un courant électrique traverse une couche ferromagnétique, les spins sont filtrés en fonction du sens de l'aimantation du matériau. La figure 1.19 illustre le mécanisme de variation de la conductance, ou de la résistance, par effet magnéto résistance. En effet, si la couche ferromagnétique a son aimantation orientée dans la même direction que les spins majoritaires, tous les électrons ayant leur spin "up" traverseront la couche ferromagnétique facilement alors que ceux qui ont leur spin

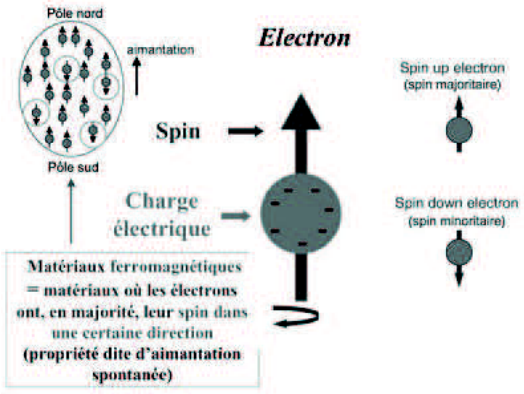


FIG. 1.18 – *Spin des électrons*

"down" seront bloqués, ralentis ou réfléchis. La résistance électrique est alors faible. A l'inverse, si la couche ferromagnétique a son aimantation dans une direction opposée aux spins majoritaires, tous les électrons ayant leur spin "up" seront bloqués ou ralentis, alors que tous les électrons ayant leurs spin "down" traverseront facilement la couche ferromagnétique. La résistance électrique est alors élevée. Les figures 1.19 (a) et (b) montre respectivement que dans une configuration où les deux couches ont leur aimantation dans le même sens (a) la résistance équivalente est faible, alors que dans le cas où leur aimantation sont dans le sens opposé la résistance équivalente est forte (b).

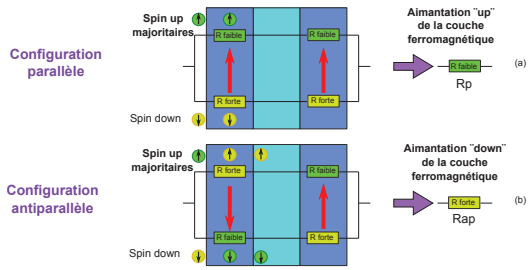


FIG. 1.19 – *Effet Magneto Résistance*

Le concept général de la spintronique est de placer des matériaux ferromagnétiques sur le trajet des électrons et d'utiliser l'influence du spin sur la mobilité des électrons dans ces matériaux. Cette influence, d'abord suggérée par Mott [92] en 1936, a été ensuite démontrée expérimentalement et décrite théoriquement à la fin des années 60 [36] [79].

Le développement de la spintronique a suivi la découverte de la magnétorésistance géante (GMR) en 1988 [12] [20]. Ces travaux ont été mené conjointement par

les équipes d'Albert Fert (Kernforschungsanlage Jülich GmbH), physicien français spécialiste de physique de la matière condensée, professeur émérite à l'Université Paris-Sud 11 en 2010 et directeur scientifique au sein de l'unité mixte de recherche CNRS/Thales, et par l'équipe de Peter Grünberg, physicien allemand, retraité depuis 2004 mais continuant à s'impliquer en tant qu'invité au Centre de recherches de Juliers. En 2007, ils reçoivent conjointement le prix Nobel de physique pour la découverte de la magnétorésistance géante. La découverte de la GMR est largement considérée comme le début de la spintronique. La GMR est une variation de résistance notable dans un empilement de couches ferromagnétiques et métalliques en fonction du champ magnétique appliqué.

La spin valve est le premier dispositif basé sur la magnéto résistance géante (GMR) exploitable pour des applications spintronique [32]. En effet un champ de quelques mT est nécessaire alors que les expériences de Fert et Grünberg nécessitaient des champs de l'ordre du Tesla. Dans une spin valve deux couches ferromagnétiques sont séparées par une couche non magnétique conductrice d'environ 3 nm, mais sans couplage RKKY. Le couplage RKKY, pour Ruderman-Kittel-Kasuya-Yosida, est une interaction quantique de couplage entre des moments magnétiques nucléaires ou des spins d'électrons localisés de la couche interne d'un métal via les électrons de conduction. Si le champ coercitif des deux électrodes ferromagnétiques de la spin valve est différent, il est alors possible d'en commuter qu'une seule des deux seulement. Ainsi, on peut réaliser un alignement parallèle ou antiparallèle entre ces couches, où la résistance est plus grande dans le cas antiparallèle. Dans ce type de structure, la principale caractéristique de cet empilement est que les 2 couches ferromagnétiques sont séparées par un métal classique et non par un isolant comme dans la TMR. La magnétorésistance géante par spin valve présente beaucoup d'intérêts industriels et commerciaux car c'est la configuration utilisée dans les têtes de lecture de disques durs.

La découverte de la GMR a conduit aux premières utilisations pratiques de cette influence. De nombreux autres phénomènes exploitant aussi le spin des électrons se sont ensuite révélés et, aujourd'hui, la spintronique se développe dans de très nombreuses directions: magnétorésistance tunnel, phénomènes de transfert de spin, spintronique avec semi-conducteurs, spintronique moléculaire, spintronique avec multiferroïques [71]. La spintronique a pour but d'augmenter les performances des composants semi-conducteurs, soit du point de vue énergétique, soit du point de vue fonctionnalité.

1.3.4.2 Magneto résistance tunnel

Il a été démontré que les électrons sont capables de traverser un isolant pour passer d'une couche conductrice à une autre couche conductrice par effet tunnel, en 1960 par I. Giaever. Il observa alors le déplacement des électrons à travers un isolant entre 2 couches, une supraconductrice et l'autre conductrice. Il a été récompensé par le prix Nobel de la Physique en 1973 [41] [4]. En 1971 l'effet de conservation de spin à travers un tel tunnel a été démontré par P.M. Tedrow et R. Meservey [115]. Le spin est une propriété quantique intrinsèque associée à chaque particule, qui est caractéristique de la nature de la particule au même titre que sa masse et sa charge électrique. L'effet tunnel magnétique a lui été démontré plus tard, pour la première fois par M. Jullière en 1975. Il observa que la conductance d'une jonction composée de 2 couches ferromagnétique et séparées par un isolant était liée à l'orientation magnétique relative des 2 couches ferromagnétiques (Fe-Ge-Co). Cet empilement typique de couches "magnétique / isolant / magnétique" est appelé jonction tunnel magnétique (JTM), composant faisant partie intégrante des mémoires MRAM aujourd'hui.

Les jonctions tunnel des mémoires MRAM sont composées de 2 couches ferromagnétiques séparées par un isolant appelé barrière tunnel, pour laquelle la barrière d'énergie est exprimée en électronVolt (eV). La figure 1.20 montre la structure de base d'une telle jonction, soit dans une configuration où les 2 aimantations sont dans le sens parallèle (a), soit dans le cas où les aimantations sont dans le sens antiparallèle (b).

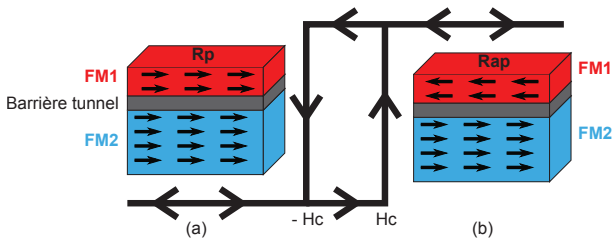


FIG. 1.20 – *Jonction Tunnel Magnétique*

Il est possible de changer la direction de l'aimantation d'une couche ferromagnétique en appliquant un champ extérieur suffisamment important. L'aimantation de la couche FM2 est choisie et imposée pendant la fabrication de la jonction tunnel, alors que l'aimantation de la couche de référence peut être changée dynamiquement par application d'un champ magnétique extérieur pendant l'utilisation du composant. Ce champ d'écriture doit être suffisamment élevé pour changer l'aimantation de la couche de stockage mais suffisamment faible pour ne pas changer en même temps

celui de la couche de référence.

Le cycle d'hystérésis représenté sur la figure 1.20 matérialise le fait qu'il est nécessaire d'appliquer un champ positif pour passer de l'état parallèle à l'état antiparallèle, et d'appliquer un champ négatif pour passer de l'état antiparallèle à l'état parallèle. La valeur absolue de ce champ est la même, classiquement appelé champ coercitif Hc, fonction de l'angle entre les aimantations des couches.

La valeur de la résistance peut s'exprimer selon l'*equation 1.4*. Rp est la résistance faible, c'est à dire la résistance lorsque les aimantations de la couche de stockage et la couche de référence sont alignées dans le même sens. θ représente l'angle entre l'aimantation de la couche de référence et l'aimantation de la couche de stockage. ΔR exprime la variation de résistance entre les états parallèle et antiparallèle.

$$R(\theta) \ = Rp + \Delta R \times \frac{(1 - cos(\theta))}{2} \tag{1.3}$$

En effet, lorsque les aimantations sont dans le même sens on a $\theta = 0$ degré et donc R = Rp. Lorsque que les aimantations sont dans le sens opposés, $\theta = 180$ degrés, et donc $R = Rp + \Delta R =$ Rap. Cette relation exprime également souvent les conductances en effet tunnel, suivant cette relation:

$$G(\theta) \ = Gp - \Delta G \times \frac{(1 - cos(\theta))}{2} \tag{1.4}$$

La variation de résistance dans les jonctions tunnel magnétiques est exprimée en pourcentage par la TMR: Tunnel Magneto Resistance, selon l'*equation 1.5* suivante:

$$TMR\ (\%) \ = \ \frac{\Delta R}{Rp} \ = \ \frac{Rap - Rp}{Rp} \ = \ \frac{\Delta G}{Gap} \ = \ \frac{Gp - Gap}{Gap} \tag{1.5}$$

Rp est la valeur de la résistance lorsque les aimantations des 2 couches magnétiques sont dans l'état parallèle et Rap est la résistance lorsque les aimantations de ces 2 mêmes couches sont dans l'état antiparallèle.

Les premières expériences, faites à température ambiante, ont montré des TMR de l'ordre de la dizaine de pourcent. Il s'agit des travaux de J. Moodera en 1995 [90]. Par la suite, beaucoup de travaux ont été faits, notamment au niveau procédé et couches magnétiques. En 2004, Parkin et Yuasa étaient capables de fabriquer des jonctions à base d'oxyde de magnésium MgO du type Fe/MgO/Fe, atteignant une TMR de 200% à température ambiante[95]. En 2009, des TMR de 600% ont été observées, toujours à température ambiante [75] et à base de MgO, mais avec des couches magnétiques à base de Fer, Colbalt et Bore (CoFeB/MgO/CoFeB). Dans une

spin valve, les résistances Rp et Rap sont plus petites que pour une jonction tunnel, ainsi que sa variation relative de résistance, et le comportement est plus linéaire. Une spin valve est plus sensible aux variations de champs magnétiques, c'est pourquoi elles prennent souvent places dans des applications de type capteurs magnétiques.

Grace aux différentes recherches et découvertes de ces dernières années, le niveau de la TMR n'a eu de cesse que d'augmenter, ce qui est une satisfaction du point de vue de la conception microélectronique. En effet, dans le but d'intégrer des composants magnétiques du type jonctions tunnel magnétiques aux procédé CMOS, toujours de plus en plus à la pointe de la technologie, et toujours de plus en plus performants, il est important d'avoir une forte TMR. Cela permet de pouvoir coder et/ou décoder facilement et de façon stable un niveau logique '1' ou '0', ce qui rend les circuits plus robustes aux variations des procédés de fabrication. Comme nous le présenterons dans le chapitre 4, le fait d'avoir une forte TMR permet d'améliorer les performances des structures différentielles, dans lesquelles 2 JTM sont utilisées systématiquement en opposition d'état, c'est à dire que l'une est dans un état parallèle alors que l'autre est dans un état antiparallèle.

La TMR dans les jonctions tunnel magnétiques peut aussi s'exprimer en fonction de la polarisation des spins des électrons, définie par la relation de Jullière ci-dessous (*equation 1.6*) [78], dans laquelle P1 est la polarisation des spins des électrons de la première couche ferromagnétique, et P2 celle de la deuxième électrode ferromagnétique.

$$TMR = \frac{2P1P2}{1 - P1P2} \tag{1.6}$$

La polarisation elle-même est définie par le rapport de densité des électrons ayant leur spin "up" (Dup) et la densité des électrons ayant leur spin "down" (Ddown) au niveau de Fermi, pour chaque couche, comme précisé sur l'*equation 1.7*.

$$P = \frac{Dup - Ddown}{Dup + Ddown} \tag{1.7}$$

Dans la mesure où la densité des spins "up" et la densité des spins "down" sont égales dans les matériaux paramagnétiques, on a P=0 pour ces matériaux. En revanche, dans les matériaux ferromagnétiques qui composent les électrodes des jonctions tunnel magnétiques, la densité des spins "up" et "down" est différente. On a donc $P > 0$ mais toujours $P < 1$. Par exemple, pour le Fer et le Cobalt, le rapport des densités de spins est de l'ordre de 3. Cela implique donc que:

$$Dup = 3.Ddown \text{ soit } P1 = \frac{3.Ddown - Ddown}{3.Ddown + Ddown} = \frac{2}{4} = 50\%$$

On peut en extraire la TMR d'une JTM composée de Cobalt ou de Fer:

$$TMR_{Co} = TMR_{Fe} = \frac{\Delta R}{Rp} = \frac{2P1P2}{1 - P1P2} \tag{1.8}$$

avec P1 = P2 car les 2 couches magnétiques sont identiques.

$$TMR_{Co} = TMR_{Fe} = \frac{2 \times 0.50 \times 0.50}{1 - 0.50 \times 0.50} = 66\% \tag{1.9}$$

La densité de polarisation de spin dans un matériau métallique de transition est illustré en figure 1.21. Soit le matériau est dit paramagnétique, c'est à dire que les moments magnétiques ne sont pas ordonnés dans le matériau et donc que la somme des moments est nulle. Soit le matériau est dit ferromagnétique, c'est à dire que les moments magnétiques sont tous ordonnées, avec des moments majoritaires (spin up) et d'autres minoritaires (spin down), et donc que la somme des moments est non nulle. Dans le cas d'un matériau dit ferromagnétique semi-métal, la densité des moments minoritaires au niveau de Fermi peut être nulle et donc la polarisation peut atteindre la valeur limite P = 1 à 0 Kelvin.

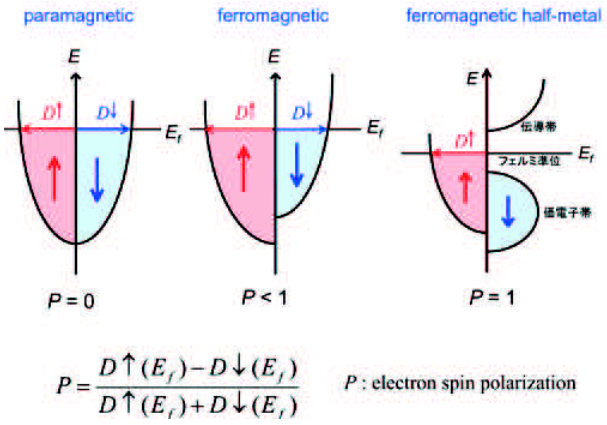


FIG. 1.21 – *Densité de polarisation de spins dans un matériau [72]*

Dans l'objectif d'avoir une TMR la plus importante possible, on peut déduire de ces équations qu'il faut que le matériau soit tel qu'il ait un contraste de densités d'états entre spin "up" et spin "down" le plus grand possible.

1.3.4.3 Fonctionnement général des MRAM

Les mémoires MRAM, pour Magnetic RAM, font parties de la catégorie de mémoires ferromagnétiques, car elles sont composées de 2 couches magnétiques séparées par un isolant. L'élément principal des mémoires MRAM est la jonction tunnel

magnétique (JTM), dont les différentes variantes sont décrites ci-après dans ce manuscrit, ainsi que leur mode de lecture et d'écriture. Ces jonctions sont classifiées en 2 catégories, les jonctions à écriture par champ magnétique externe et les jonctions à écriture par courant polarisé en spin. Les mémoires MRAM sont aujourd'hui considérées comme de très bonnes candidates parmi toutes les mémoires non volatiles émergentes, grâce à leurs caractéristiques et propriétés intéressantes pour une majeure partie des applications du monde actuel: vitesse, consommation, endurance quasi infinie, immunité aux radiations et miniaturisation par exemple [26].

1.3.4.4 Ecriture FIMS: Field Induced Magnetic Switching

Les jonctions tunnel magnétiques ont été intégrées aux procédés de fabrication CMOS au cours des 10 dernières années. Dans le cas de JTM classiques présentées sur la figure 1.20, le champ coercitif nécessaire au changement de l'aimantation de la couche de stockage est si élevé, autour des 90 Oe, que le courant nécessaire pour générer ce champ est bien trop grand. Dans le but d'intégrer les JTM dans des mémoires de technologies CMOS, la première génération de JTM FIMS utilise 2 champs magnétiques perpendiculaires appliqués simultanément permettant de sélectionner précisément un point mémoire à l'intersection d'une ligne et d'une colonne d'une matrice mémoire. Le principe de fonctionnement est illustré sur la figure 1.22.

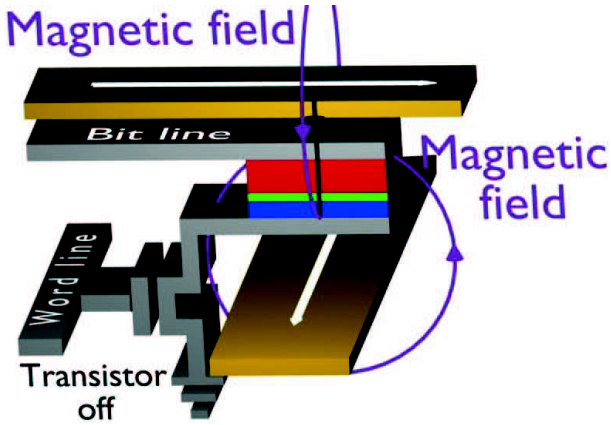


FIG. 1.22 – *Schéma de principe FIMS*

Chaque bit mémoire est composé d'une jonction tunnel magnétique, connectée d'un coté à un transistor de commande lui-même piloté par le signal Word Line (WL) et de l'autre à une ligne appelée Bit Line (BL). Au-dessus de la Bit Line passe une première ligne de champ d'écriture. Au-dessous de la jonction passe une seconde ligne de champ d'écriture. Notons que dans certaines architectures, la ligne

de champ d'écriture la plus au-dessus peut être supprimée. Dans ce cas, la bit line sert aussi de ligne de champ d'écriture. Le fait d'utiliser 2 lignes de champ permet de mutualiser entre plusieurs jonctions les 2 lignes d'écriture dans une approche de type mémoire dont le schéma de principe est illustré sur la figure 1.23. En effet, une word line sélectionne un ensemble de JTM sur 1 colonne et une bit line sélectionne un ensemble de JTM sur une ligne. Par exemple, la colonne WL5 sélectionne plusieurs JTM mais sans générer assez de champ pour les écrire, il s'agit des JTM en jaune. La ligne BL2 sélectionne d'autres JTM, de même sans générer assez de champ pour les écrire. Il s'agit des JTM en bleu. Seule 1 jonction tunnel sera sélectionnée par ces 2 connexions WL5 et BL2, soit la JTM verte. L'association de ces 2 champs étant suffisante pour retourner l'aimantation de la couche de référence de cette jonction permet donc une écriture d'un bit d'une mémoire.

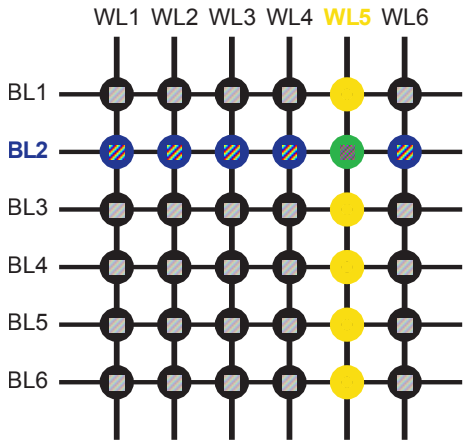


FIG. 1.23 – *Réseau de points mémoires écrits par la méthode FIMS*

En revanche, l'association de ces 2 champs d'écriture doit être suffisamment élevée mais sans être surdimensionnée, au risque d'écrire les jonctions voisines soumises à un seul champ, ce qui aurait pour conséquence de rendre le système non fonctionnel. Il est donc important de contrôler et maitriser l'intensité du champ généré pour l'écriture des JTM. Le théorème de Stoner Wohlfarth [110] exprime à travers un astéroïde présentée en figure 1.24, la combinaison de champs nécessaires pour changer l'aimantation d'un matériau ferromagnétique en fonction de l'amplitude des 2 champs respectivement appliqués parallèlement à l'axe facile de la jonction.

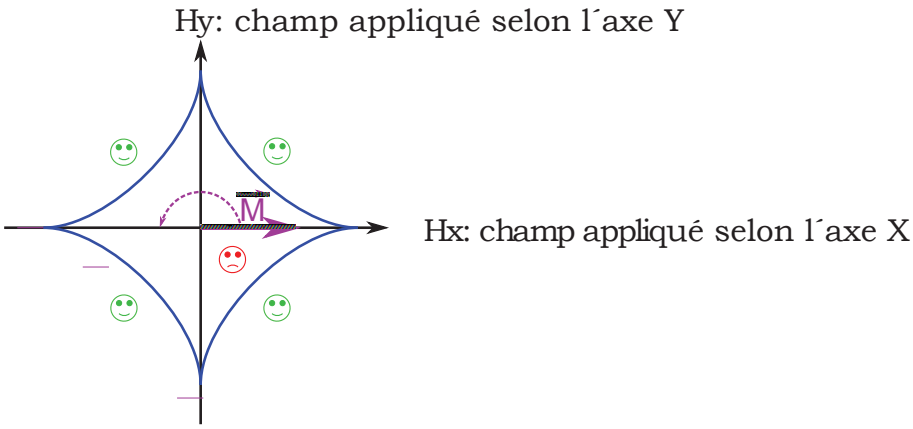


FIG. 1.24 – *Astéroïde de Stoner - Wohlfarth*

Prenons l'exemple de la figure 1.24 où l'aimantation $\vec{M}$ est orientée vers la droite selon l'axe des X positifs. Si on applique un champ supérieur à 90 Oe à la fois suivant les axes X et Y, toutes les jonctions sur une ligne ou une colonne sont écrites, ce qui n'est pas ce que l'on souhaite. Si on applique un champ inférieur à 30 Oe suivant les axes X et Y, aucune jonction ne sera écrite, ce qui n'est pas ce que l'on souhaite non plus. En revanche, si on applique un champ de valeur intermédiaire, typiquement 50 à 60 Oe suivant les 2 axes X et Y, on écrit seulement la jonction qui voit la somme des 2 champs. La marge d'écriture est de l'ordre de $+/-$ 20 Oe. Cependant, avec la dispersion des propriétés des jonctions, notamment géométriques et de fluctuations thermiques d'aimantation, l'astéroïde est différent d'une jonction à l'autre, réduisant cette fenêtre d'application du champ. Ceci est de plus en plus vrai lorsque la taille diminue car les variations relatives de procédé augmentent. Ainsi dans un but ultime d'intégration de JTM dans l'électronique CMOS et de miniaturisation, on s'aperçoit que cette méthode a ses limites. Lorsque l'on réduit la taille de la jonction, certes le champ nécessaire diminue également car il dépend du volume de la couche à écrire, mais les variations de procédé de fabrication deviennent de plus en plus prépondérantes. Cela revient à dire que l'astéroïde ne présente plus de symétrie, ni par rapport à l'axe des X, ni par rapport à l'axe des Y, ni par rapport au centre, et surtout qu'il n'est plus le même pour toutes les jonctions voisines. Ces variations sont telles pour des composants très petits, que la géométrie de chaque composant, les JTM en l'occurrence ici, n'est pas assez déterministe. Si on ajoute à cela que plus les technologies CMOS avancent plus les transistors sont petits, donc plus la surface occupée par une jonction est petite, alors plus les jonctions sont proches les unes des autres. La densité de jonction tunnel augmente par unité de surface. On se confronte alors très vite à des problèmes de sélectivité avec une telle

méthode d'écriture, car le risque de générer de façon certaine un champ d'écriture suffisant pour une JTM entraine le risque fort d'écrire une jonction voisine. C'est pourquoi la société Everspin a mis au point une variante de cette méthode pour palier partiellement à ce problème, expliquée dans le prochain paragraphe.

1.3.4.5 Ecriture Toggle - FIMS

De la même façon que dans la méthode d'écriture FIMS, la méthode Toggle utilise des jonctions tunnel magnétiques. La figure 1.25 montre une structure de la MRAM TAS et son empilement [33]. Elle comporte un élément de type JTM connecté à l'électronique sur ses 2 terminaux T1 et T2, ainsi que 2 lignes d'écriture, word line et bit line. Celles-ci sont utilisées pour générer les champs magnétiques d'écriture par un courant électrique. L'élément magnétorésistif est composé de 5 couches: une couche ferromagnétique pour laquelle l'aimantation est figée et sert de couche de référence (CR). Une barrière tunnel permettant l'effet de magnétorésistance, au-dessus de laquelle se trouve une couche anti-ferromagnétique de synthèse (AFS). Celle-ci est elle-même composée de 2 autres couches ferromagnétiques libres dont les aimantations sont couplées antiparallèlement l'une à l'autre, servant de couches de stockage (CS1 et CS2), séparées par une couche de ruthénium dont la fonction est d'exercer ce couplage antiparallèle.

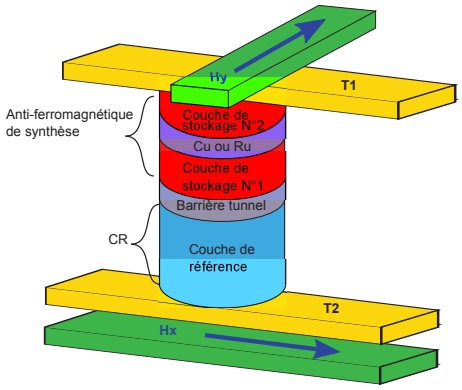


FIG. 1.25 – *Structure d'une Jonction Tunnel Magnétique Toggle*

Ce type d'empilement AFS a un volume magnétique global plus important, ce qui apporte une augmentation de la stabilité de l'information vis à vis des fluctuations thermiques, et réduit le couplage magnéto-statique entre couche de stockage et couche de référence. Ces 2 couches de stockage ont leur aimantation en opposition, soit à 180° l'une par rapport à l'autre. Dans leurs états stables, l'une a son aimantation orientée à 45° et l'autre à 225°. Le principe d'écriture est alors d'inverser l'aimantation de

ces 2 couches de stockage. Cette inversion se fait par application d'une séquence de champs magnétiques perpendiculaires sur 2 lignes que l'on nommera bit line et word line. Cette séquence correspond à une succession de rotation de champ global appliqué par pas de 45°. Ce type de cellule a remplacé la précédente car elle est moins sensible au problème de sélectivité, du fait que la méthode applique 2 champs non simultanés selon une séquence bien spécifique. Ces problèmes de sélectivité, qui impliquent le risque d'écrire une jonction voisine en appliquant un champ soit sur la word line soit sur la bit line sont décrits dans [120] et [121]. La résistance de ce type de jonction est déterminée par la direction de l'aimantation de la couche de référence et celle de la couche de stockage au contact de la barrière tunnel.

L'écriture de cet élément consiste donc à imposer la direction de l'aimantation des 2 couches de stockage, selon une séquence très spécifique, comme il est décrit dans [103] et [33]. Bien qu'il existe plusieurs séquences aboutissant au même résultat, la plus courante "box field" est présenté en figure 1.26.

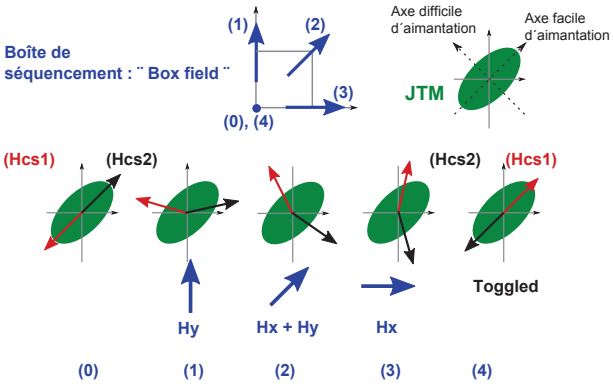


FIG. 1.26 – *Séquencement d'écriture par la méthode Toggle MRAM*

La première phase d'écriture consiste à faire circuler un courant dans la ligne word line de façon à avoir un champ appliqué selon l'axe Y (1). Les 2 aimantations tendent à s'aligner de façon perpendiculaire à ce champ Hy. Ensuite, un autre champ est appliqué en même temps dans la ligne "bit line", selon l'axe X (2). L'aimantation de chaque couche reste dans le même état l'une par rapport à l'autre, mais changent de direction pour rester perpendiculaires au champ total appliqué, Hx + Hy. Enfin, seul le champ selon l'axe X est maintenu (3). A nouveau, les 2 aimantations changent de direction pour rester perpendiculaires au champ appliqué Hx. Lorsque ce dernier champ est arrêté, les 2 aimantations basculent dans leur état stable le plus proche (4), celui qui nécessite le moins d'énergie pour se stabiliser. La forme elliptique de la jonction tunnel fait que les aimantations se stabilisent selon l'anisotropie, de la même façon que dans l'état stable avant écriture, mais chacune avec leur aimantation se

trouvant alors dans une direction opposée.

Aujourd'hui, la société Everspin qui a mis en place ce procédé d'écriture Toggle, est la seule à commercialiser des produits industriels MRAM. Elle propose des mémoires de 256 Kb à 16 Mb, port série ou 8/16 bits, dans une gamme de température de 0°C à 125°C [66]. Everspin a vendu plus de 2.5 millions de mémoires en 2011 et prévoit d'en vendre près de 5 millions en 2012, comme le montre leurs prévisions sur la figure 1.27 [65]. Néanmoins, cette société annonce l'arrivée de leur nouvelle génération de MRAM prochainement, basée sur la méthode d'écriture STT.

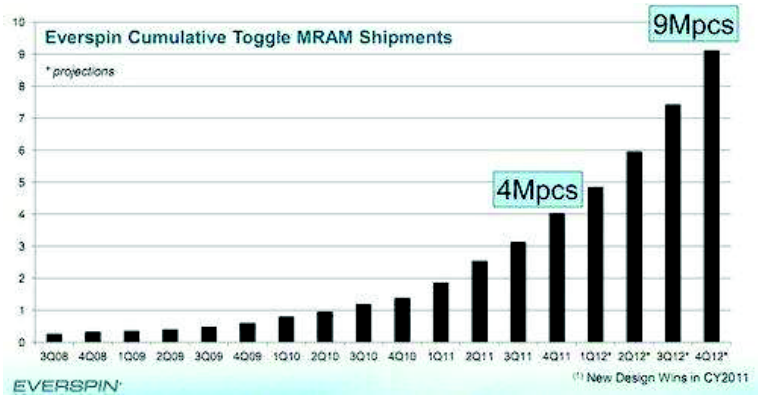


FIG. 1.27 – *Prévision des ventes de Everspin*

1.3.4.6 Ecriture TAS: Thermally Assisted Switching

Malgré la solution Toggle proposée par Everspin, la MRAM ne semblait pas avoir beaucoup d'avenir dans les technologies très submicroniques au regard de certaines de ses concurrentes. Le principal problème dans la miniaturisation de ce type de composant est la sélectivité, en plus de la consommation, car le risque d'écriture des éléments voisins est fort. Une nouvelle méthode d'écriture des JTM a été développée et brevetée par le laboratoire CEA/CNRS SPINTEC, en 2001. Par la suite la société Crocus Technology a été créée afin de commercialiser la mémoire dites TAS-MRAM. Les JTM FIMS étant composée de 2 couches ferromagnétiques ou intégrant un niveau anti-ferromagnétique de synthèse pour le Toggle FIMS, Crocus Technology propose une jonction pour laquelle la stabilité de l'aimantation des 2 couches ferromagnétiques est assurée par couplage avec deux couches anti-ferromagnétiques, grâce au phénomène dit d'anisotropie d'échange. Une particularité d'un matériau anti-ferromagnétique est que son aimantation globale est nulle car la somme des moments magnétiques est nulle. Contrairement aux matériaux paramagnétiques qui ont aussi une aimantation globale nulle, les moments magnétiques d'un matériau

anti-ferromagnétique sont ordonnés dans une direction donnée ou dans une direction opposée à celle-ci, une rangée sur 2 dans l'épaisseur. Une jonction TAS étant ronde, contrairement à une jonction FIMS qui est elliptique, la stabilité est assurée par le fait qu'à l'interface entre la couche ferromagnétique et la couche anti-ferromagnétique, les spins s'alignent sur ceux de la couche de stockage. Les 2 couches anti-ferromagnétiques étant intrinsèquement stables à température ambiante, elles maintiennent et imposent une aimantation chacune à leur couche voisine, dans le sens de l'aimantation des moments à leur interface. L'une d'elle est placée en dessous de la couche de référence, selon une direction choisie lors de la fabrication. L'autre est placée au-dessus de la couche de stockage. Chacune de ces couches a une température de blocage différente, ce qui permet lors de l'écriture de libérer l'aimantation de la couche de stockage uniquement et ainsi autoriser le changement de son aimantation, tout en gardant piégée l'aimantation de la couche de référence. La figure 1.28 montre ce type d'empilement, ainsi que l'orientation des spins de chaque matériau, où l'on retrouve l'ensemble des couches décrites précédemment, ainsi que la barrière tunnel MgO (Magnesium Oxide) entre les 2 couches ferromagnétiques. Notons qu'une jonction TAS peut aussi intégrer une barrière thermique qui permet de concentrer la chaleur à travers les matériaux, non représentée sur cette figure.

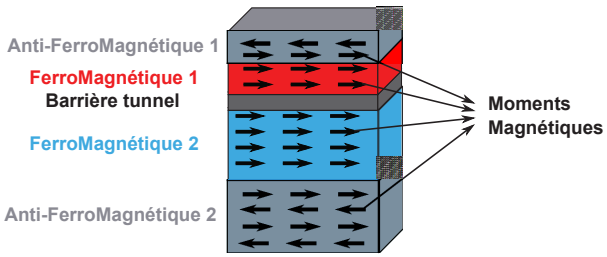


FIG. 1.28 – *Empilement d'une jonction TAS*

La particularité de ce type d'empilement est qu'il est très stable du point de vue de l'aimantation, à température ambiante. Afin de pouvoir changer l'aimantation de la couche de référence il est nécessaire de chauffer la couche anti-ferromagnétique placée au-dessus de la couche de référence, au-delà de sa température de blocage, soit environ 150°C [14], alors que la température de blocage de la couche anti ferromagnétique sous la couche de référence est d'environ 340°C. On dit que l'écriture est assistée thermiquement. Lorsque la jonction est suffisamment chauffée, les spins de la couche de stabilisation anti-ferromagnétiques sont complètement désordonnés. Seul un faible champ magnétique est nécessaire pour imposer une aimantation à la couche de stockage, pour laquelle les spins vont être dans une direction dépendante

du champ. La température ne doit pas non plus être excessive pour ne pas risquer de dépiéger l'aimantation de la couche de référence. Après un certain temps, quelques nanosecondes, la jonction n'est plus chauffée car aucun courant électrique ne la traverse. C'est donc la phase de refroidissement. Le champ doit être maintenu pendant toute la durée de refroidissement pour s'assurer de l'écriture et éviter que l'aimantation ne change pendant cette phase. On parle alors de refroidissement sous champ. Lorsque la température de la jonction est de nouveau en dessous de la température de blocage de la couche anti-ferromagnétique, les spins de cette dernière se réordonnent, toujours dans le même sens que ceux qui sont à l'interface des 2 couches, une rangée sur 2 comme précédemment. La jonction est alors dans un nouvel état stable d'aimantation, à température ambiante. Le fonctionnement de cette technologie TAS est illustré par le résultat de simulation figure 1.29.

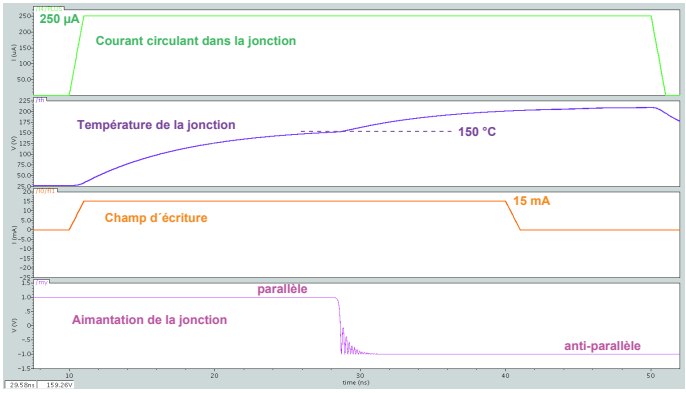


FIG. 1.29 – *Simulation d'une JTM selon la méthode TAS*

On remarque que dès qu'un courant circule à travers la jonction, la température de celle-ci s'élève. Lorsque qu'un champ d'écriture est appliqué et que la température a atteind la température de blockage, l'aimantation peut alors changer d'état. Le champ d'écriture peut être appliqué en phase terminale du chauffage seulement.

Dans les circuits où les jonctions sont organisées de façon matricielle, comme c'est le cas dans une architecture mémoire par exemple, les lignes de champs d'écriture ainsi que les générateurs de courant nécessaires à la génération de ce champ sont mutualisés entre plusieurs jonctions. Cette méthode TAS permet de palier au problème de sélectivité, car en effet il est facile de maitriser les jonctions qui seront chauffées et celles qui ne le seront pas, dans la mesure où cela consiste à contrôler un courant électrique. Le principe est le même que pour l'adressage de n'importe quel type de mémoire, utilisant 2 signaux "bit line" et "word line" par point mémoire.

La méthode d'écriture TAS, propre à la société Crocus Technology, a néanmoins

une double contrainte à gérer, le chauffage des jonctions et le champ d'écriture. Cette écriture se fait en 2 temps dans les applications de type circuit logique car ils sont la plupart du temps composés de 2 JTM en opposition d'aimantation. La ligne de champ d'écriture étant la même pour les 2 JTM et pour plusieurs structures, il faut générer le courant une fois dans un sens et une fois dans le sens opposé pour écrire successivement toutes les jonctions qui doivent être écrites avec une aimantation parallèle, puis toutes celles qui doivent être écrites avec une aimantation antiparallèle. Cette méthode est par conséquent moins gourmande en énergie que la méthode FIMS car elle ne nécessite qu'un seul champ d'écriture contre 2 pour la méthode FIMS. De plus, le champ d'écriture nécessaire est beaucoup plus faible. Enfin, la méthode TAS est relativement miniaturisable car le courant nécessaire au chauffage diminue fortement lorsque la taille de la jonction diminue, alors que le champ d'écriture ne dépend pas de la taille mais de la géométrie physique.

1.3.4.7 Ecriture STT: Spin Transfer Torque

Comme nous venons de le préciser précédemment, il existe aujourd'hui plusieurs types de jonctions tunnel magnétiques, chacune ayant ses limitations. La méthode FIMS est limitée en termes de miniaturisation à cause des champs nécessaires importants mais également parce que ces champs ne diminuent pas lorsque la taille de la jonction diminue. La méthode TAS est limitée par sa consommation, car en effet même si le chauffage ne requiert pas beaucoup d'énergie, le courant nécessaire pour générer le champ d'écriture reste relativement élevé, bien que plus faible qu'avec les méthodes FIMS, et ne diminue pas non plus lorsque la taille de la jonction diminue. L'énergie totale reste tout de même beaucoup plus élevée que l'énergie nécessaire pour changer l'aimantation de la couche de référence par la méthode STT, Spin Transfer Torque.

Les premiers travaux de recherche sur la méthode d'écriture STT ont commencé dans les années 70 et 80 avec les prédictions de Berger sur le fait que l'effet de transfert de spin pourrait déplacer les parois de domaine magnétique [16], suivi par les expérimentations de son groupe sur l'observation du déplacement des parois de domaine dans les matériaux ferromagnétiques sous l'influence d'impulsions de courant importantes [17] [37]. A ce moment-là, l'enthousiasme n'était pas franchement réel, principalement à cause du fait que les courants nécessaires étaient exorbitants, jusqu'à 45 ampères [98]. Le déclenchement des travaux de recherche sur le spin transfer torque s'est fait suite aux 2 publications de Slonczewski [109] et de Berger [18] qui ont prédit indépendamment que la circulation d'un courant suffisamment fort perpendiculaire au plan dans un empilement métallique peut réorienter l'aimantation

d'une couche de cet empilement. Deux à trois années plus tard, l'effet fut démontré expérimentalement par différentes équipes scientifiques à travers le monde, et cela dans différentes configurations de structures nanométriques : à Grenoble par l'équipe de Tsoi et al. [116], à Cornell par l'équipe de Ralph et al. [93] et à Paris par l'équipe de Fert et al. [42], chacune dans des structures sandwich de Co/Cu/Co.

Ce type de mémoire cumule beaucoup d'avantages, à savoir une faible consommation car elle ne nécessite que de 100 à 200 μA en technologie 90nm et bien moins de 100 μA en technologie 45 nm et en dessous. De plus, son temps d'écriture et de lecture est très rapide, quelques nanosecondes seulement, sa densité est proche de celle des DRAM, de 6 à 20 F^2. Son endurance est proche de la SRAM c'est à dire autour des 10^{16} cycles et bien entendu sa non volatilité est un élément clef. Cependant, le procédé de fabrication reste encore très pointu et relativement compliqué, avec pas moins de 10 à 12 niveaux physiques, pour une épaisseur de 0.8 à 2 nm [69]. Cette technologie, compatible avec les technologies CMOS, pourrait éventuellement dans les années à venir remplacer la plupart des mémoires utilisées dans les applications actuelles.

La méthode d'écriture STT des jonctions tunnel magnétiques, aussi appelée CIMS pour Current Induced Magnetic Switching, reprend les principes fondamentaux de la spintronique décrit dans le paragraphe 2.3.4.2 "La spintronique". Le principe est d'imposer une aimantation à la couche de référence en fonction du sens du courant polarisé en spin qui la traverse. L'empilement classique d'une jonction tunnel magnétique STT intègre un polariseur de spin, qui est en général la couche de référence elle-même. Ce n'est pas le cas par exemple pour un polariseur perpendiculaire permettant de faire du retournement précessionnel pour des oscillateurs, entre autres. Dans tous les cas, sa polarisation dépend de son aimantation, qui est imposée lors de la fabrication. Dans un courant électrique, la polarisation des spins est quelconque. En revanche, lorsque ce même courant traverse un matériau ferromagnétique, il acquiert une polarisation. Une jonction tunnel telle que présentée ici n'a qu'un seul polariseur. La figure 1.30 décrit le mécanisme de retournement de l'aimantation de la couche de stockage par transfert de spins imposé par un courant polarisé en spin grâce, à la couche de référence servant de polariseur. Le mécanisme est le suivant: lorsque les électrons traversent une couche ferromagnétique ils acquierent la polarisation de cette couche. Les spins majoritaires peuvent donc la traverser facilement alors que les spins minoritaires sont réfléchis.

Selon le sens du courant, celui-ci va soit traverser la jonction en passant en premier par le polariseur, soit en passant en premier par la couche de référence. La jonction étant soit dans un état parallèle soit dans un état antiparallèle, il y a donc 4 cas de

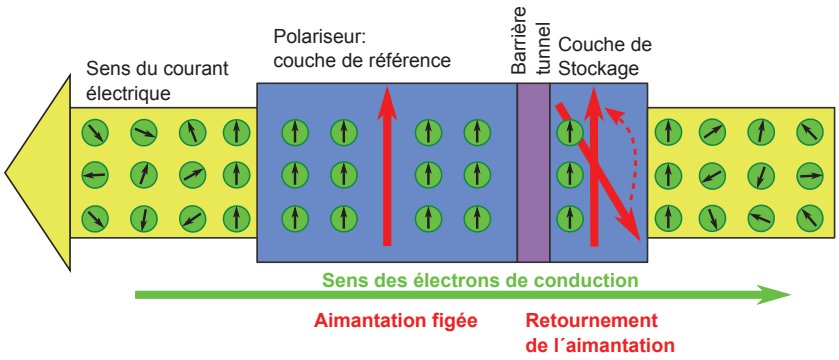


FIG. 1.30 – *Mécanisme de transfert de spin par courant polarisé en spin*

fonctionnement, décrits ci-après et illustrés sur la figure 1.31.

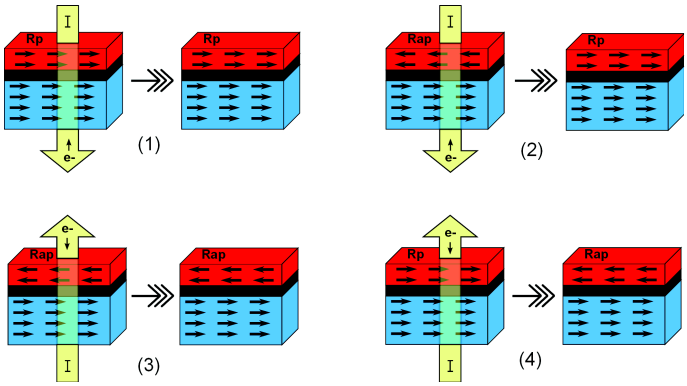


FIG. 1.31 – *Mécanisme d'écriture d'une JTM STT-MRAM*

– La jonction tunnel est dans l'état parallèle et le courant la traverse en passant d'abord par la couche de stockage: **les électrons traversent alors la JTM en passant en premier par le polariseur**. Les spins majoritaires traversent facilement la JTM alors que les spins minoritaires sont réfléchis hors de la jonction. L'aimantation de la couche de stockage est inchangée (1) car les spins la traversant sont dans le même sens.
– La jonction tunnel est dans l'état antiparallèle et le courant la traverse en passant d'abord par la couche de stockage: **les électrons traversent alors la JTM en passant en premier par le polariseur**. Les spins majoritaires traversent facilement la JTM alors que les spins minoritaires sont réfléchis hors de la jonction. L'aimantation de la couche de stockage est alors inversée (2) car les spins la traversant sont dans le sens opposé.
– La jonction tunnel est dans l'état antiparallèle et le courant la traverse en pas-

sant d'abord par le polariseur: **les électrons traversent alors la JTM en passant en premier par la couche de stockage**. Les spins majoritaires sont réfléchis par la couche de stockage hors de la jonction et les spins minoritaires sont réfléchis par la couche de référence à travers la couche de stockage. L'aimantation de la couche de stockage est inchangée (3) car les spins minoritaires sont dans le même sens que celle-ci.

– La jonction tunnel est dans l'état parallèle et le courant la traverse en passant d'abord par le polariseur: **les électrons traversent alors la JTM en passant en premier par la couche de stockage**. Les spins majoritaires traversent facilement la JTM alors que les spins minoritaires sont polarisés dans le sens opposé et réfléchis par la couche de référence à travers la couche de stockage. L'aimantation de la couche de stockage est alors inversée (4) car les spins minoritaires sont dans le sens opposé à celle-ci.

Ce dernier cas est illustré en détail sur la figure 1.32

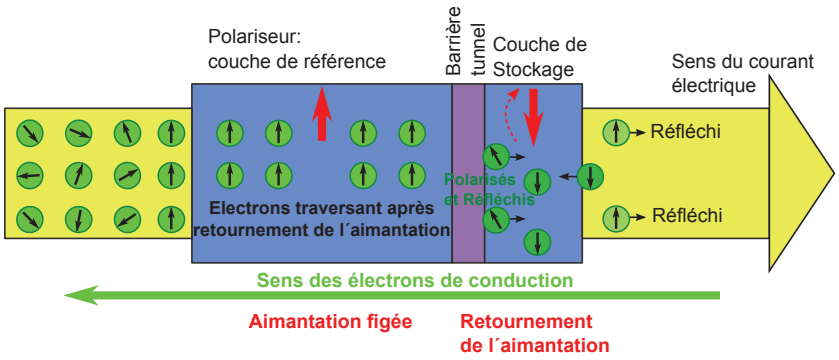


FIG. 1.32 – *Ecriture d'une jonction tunnel STT par courant polarisé en spin.*

Du point de vue de la conception électronique, on peut dire que l'état magnétique parallèle ou antiparallèle d'une jonction tunnel STT dépend, au delà d'un certain seuil de densité de courant, uniquement du sens du courant qui la traverse et qu'il suffit de générer le courant soit dans un sens soit dans l'autre pour changer la valeur de la résistance de la jonction.

Dans le cas où les spins majoritaires traversent facilement la jonction, la densité de courant nécessaire au retournement de l'aimantation de la couche de stockage est plus faible que dans le cas où les spins majoritaires sont réfléchis. Le courant électrique nécessaire pour écrire un '0' ou un '1' logique n'est donc pas tout à fait le même. Aujourd'hui beaucoup d'efforts et de travaux de recherche sont faits sur cette méthode d'écriture STT car elle semble être une réelle concurrente à la fois des mémoires non volatiles et à la fois des mémoires denses, ainsi que des mémoires

rapides du marché actuel [112]. Le premier prototype de mémoire magnétique basée sur le transfert de spin, la spin-RAM, a été présenté par Sony en 2005 à la conférence IEDM [45].

1.3.5 Synthèse sur les mémoires émergentes

Nous avons vu que toutes les mémoires non volatiles ROM, programmables ou non, ont progressivement été remplacées par d'autres mémoires, Flash majoritairement. Ces ROM ont été très utilisées depuis très longtemps mais ne sont pas les mémoires de demain, ni en termes de performances, ni en termes de besoin, ni en termes de coût. L'endurance et la vitesse sont deux inconvénients majeurs des mémoires Flash, qui occupent aujourd'hui une part de marché importante. La SRAM semble aussi avoir un avenir relativement difficile du fait que les courants de fuite deviennent très importants à partir des technologies 45n et en dessous.

En ce qui concerne les technologies émergentes, les mémoires MRAMs sont des mémoires magnétiques ayant comme propriétés d'être non volatiles, intrinsèquement insensibles aux radiations ce qui les place dans les bonnes candidates aux applications militaires et spatiales par exemple, en concurrence avec les PCRAM. De plus les MRAM sont des mémoires denses comme les DRAM, plus rapides en écriture et en lecture que la plupart de ses concurrentes non volatiles, comme les mémoires Flash et PCRAM par exemple. De plus leur extrême endurance et leur faible consommation des dernières générations leur confère de réelles qualités pour prendre une place importante dans le domaine du stockage de l'information dans toutes les applications actuelles de plus en plus gourmandes en ressources, performances tout en ayant une consommation faible.

Le tableau 1.1 fait une synthèse des avantages et inconvénients principaux des mémoires qui sont les plus utilisées dans les systèmes électroniques. Il permet également de les comparer avec la plupart des technologies émergentes [55] [69] [66], au moins avec celles pour qui les spécialistes voient un éventuel avenir dans les systèmes de demain. Les paramètres de comparaison de ce tableau sont la densité, le temps de lecture, le temps d'écriture, la consommation, la non volatilité, la miniaturisation, la facilité de mise en oeuvre pour la fabrication ainsi que l'endurance.

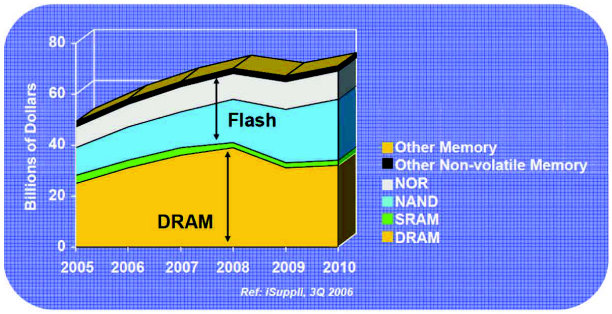
	D.(F^2)	R. (ns)	W. (ns)	Conso	NV	Min.	Fab.	End.
DRAM	6-10	10-30	15-50	Faible	Non	Oui	+	10^{16}
SRAM	50-120	1-50	1-50	Faible	Non	Oui	+	10^{16}
FLASH Nor	10	3-15	1u-10m	Forte	Oui	Non	-	10^{5}
FLASH Nand	2-5	50	0.1m-1m	Forte	Oui	Non	-	10^{3}-10^{7}
FeRAM	15-35	20-80	50	<Moyen	Oui	?	=	10^{12}
PCRAM	6-12	20-50	20-120	Faible	Oui	Moyen	=	10^{8}-10^{12}
RedOx	6-10	10-50	10-50	Faible	Oui	Oui	+	10^{7}-10^{8}
FIMS/TOG	16-40	1-20	1-20	>Moyen	Oui	Non	=	$>10^{15}$
TAS	16-40	3-20	3-20	<Moyen	Oui	<Moyen	=	$>10^{15}$
STT	6-20	2-20	2-20	Faible	Oui	Oui	-	$>10^{15}$

TAB. 1.1 – *Comparatif des mémoires*

1.4 Conclusion

De la même façon que le tableau 1.1 présenté dans la section précédente le montre, le résultat de l'étude menée par les experts de l'ITRS (International Technology Roadmap for Semiconductor's) des groupes ERD(Emerging Research Devices) et ERM (Emerging Research Materials) est que les 2 technologies STT-MRAM et RedOx-RRAM sont les technologies émergentes les plus prometteuses pour l'avenir. Ces experts pensent qu'il s'agit des technologies sur lesquelles les recherches accélérées et les développements sont recommandés. Elles ont le plus de capacités à être miniaturisées et commercialisées parmi les RAM non volatiles, et cela également audessous des générations de technologies 16 nm [69], ce qui est un élément important vu la vitesse à laquelle évoluent toujours les technologies CMOS.

Le graphique de la figure 1.33 montre clairement que parmi les mémoires de stockage de masse telles que les DRAM et les Flash, la concurrence est rude par rapport aux mémoires non volatiles émergentes, telles que les MRAM, RedOx-RRAM. Si les progrès technologiques sont dans les prochaines années à la hauteur des attentes et des espérances d'aujourd'hui, il est probable que l'une d'entre elle prenne une part importante du marché des mémoires de stockage entre autres. Il n'est donc pas impossible que dans les toutes prochaines années nous retrouvions dans nos PC portables, téléphones mobiles, PDA ou autres appareils électroniques qui font partie intégrante de notre vie quotidienne ces mémoires tant prometteuses.



Source: Gary Bronner (Rambus), Stanford EE 309 lecture, Fall 2007

FIG. 1.33 – *Marché des mémoires*

Chapitre 2

Flot de conception d'un ASIC

2.1 Introduction

L'électronique est un domaine très vaste. Depuis les premières liaisons radio du XIXème siècle aux ordinateurs portables et autres tablettes tactiles récentes, en passant par la télévision apparue dans les années 1920, par les premiers équipements automobiles de sécurité et téléphones portables, l'électronique n'a eu de cesse de progresser durant plus d'un siècle. Au fil du temps, les techniques de conception et de fabrication ont évolué, les besoins ont été de plus en plus nombreux et précis, ce qui a amené au développement de plusieurs types de circuits intégrés. Aujourd'hui, dans les systèmes électroniques en général et plus spécifiquement dans le domaine de la microélectronique, il existe 2 grandes familles de circuits intégrés. D'une part les circuits logiques programmables appelés FPGA, Field Programmable Gate Array, et les circuits spécialisés dédiés à une utilisation et à une application particulière appelés ASIC, Application Specific Integrated Circuit.

Le FPGA a pour principal avantage d'être programmable et reprogrammable après fabrication, en fonction du besoin, de l'application, de l'environnement etc. Le principe de reconfiguration remonte aux années 1960. G. Estrin proposa un processeur standard couplé à une partie matérielle reconfigurable, contrôllée par le processeur principal [34][35]. On pourrait dire qu'il s'agit là de l'ancêtre du FPGA. Un FPGA est composé de nombreuses cellules logiques élémentaires librement assemblables. Celles-ci sont connectées de manière définitive ou réversible par programmation, afin de réaliser la ou les fonctions numériques souhaitées. L'intérêt d'un tel circuit est qu'une même puce peut être utilisée dans de nombreux systèmes électroniques différents, et que sa fonction peut être modifiée à la demande de l'environnement dans lequel ce circuit évolue. La plupart des FPGA modernes sont fondés sur des cellules SRAM pour programmer aussi bien le routage du circuit que les blocs

logiques à interconnecter. Un bloc logique est de manière générale constitué d'une table de correspondance communément appelée LUT ou Look-Up-Table, dont une représentation est montrée en figure 2.1, et d'une bascule de type flip-flop [52].

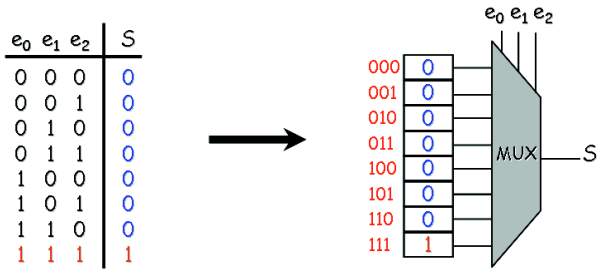


FIG. 2.1 – *Look-Up-Table*

Une LUT sert à implémenter des équations logiques, ayant généralement 4 à 6 entrées et une sortie. Elle est constituée d'une petite mémoire, d'un multiplexeur ou d'un registre à décalage. Le registre permet de mémoriser un état d'une machine séquentielle ou de synchroniser un signal. Les blocs logiques, présents en grand nombre sur la puce, de quelques milliers à quelques millions en 2007, sont connectés entre eux par une matrice de routage configurable, comme illustré en figure 2.2. Ceci permet la reconfiguration à volonté du composant, mais occupe une place importante sur le silicium et justifie le coût élevé des composants FPGA [52] pour de faibles volumes.

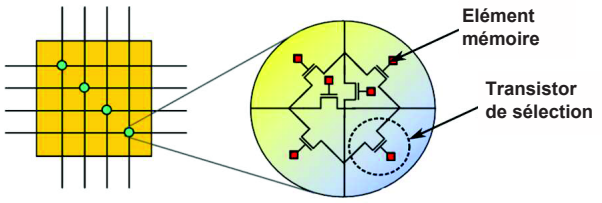


FIG. 2.2 – *Interconnexions d'un FPGA*

Les densités actuelles des FPGA ne permettent plus un routage manuel, c'est donc un outil de placement-routage automatique qui fait correspondre le schéma logique voulu par le concepteur et les ressources matérielles de la puce. Comme les temps de propagation dépendent de la longueur des liaisons entre cellules logiques, et que les algorithmes d'optimisation des placeurs-routeurs ne sont pas déterministes, alors les performances obtenues en termes de fréquence maximum dans un FPGA sont variables d'un design à l'autre. L'utilisation des ressources est par contre très bonne, et des taux d'occupation des blocs logiques supérieurs à 90% sont possibles. Le routage et les LUTs font partie de la configuration d'un FPGA et sont consti-

tués de points mémoires volatils. Il est donc nécessaire de sauvegarder le design du FPGA dans une mémoire non volatile externe, généralement une mémoire Flash série. La figure 2.3 montre une structure classique d'un FPGA volatil, qui sont les plus répandus à ce jour.

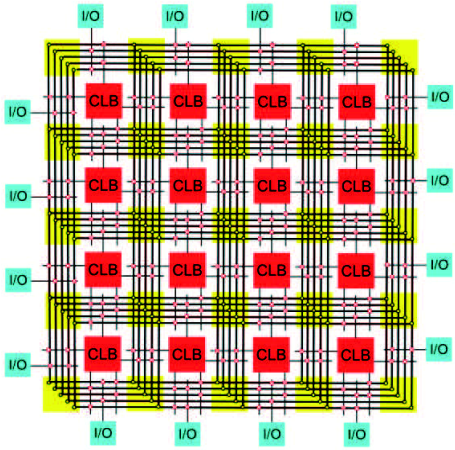


FIG. 2.3 – *Exemple de FPGA*

Certains fabricants se distinguent toutefois par l'utilisation de cellules EEPROM pour la configuration, éliminant le recours à une mémoire externe. Par ailleurs, plusieurs travaux ont été menés pour intégrer des mémoires MRAMs dans les FPGA [124], voire d'intégrer la configuration dans des registres non volatils [126].

A l'opposé, l'ASIC a pour principal avantage d'être dédié à une application. Il est développé et conçu spécifiquement pour une application selon un cahier des charges très précis. De ce fait, il peut être tout à fait optimisé, aussi bien en surface car toute la surface occupée est utilisée pour la fonction définie, qu'en vitesse. En effet les outils de conception permettent d'affiner et d'optimiser les résultats. De plus il est possible de placer les différents blocs les uns à côté des autres stratégiquement en fonction des performances attendues. Il y a aujourd'hui une multitude de procédés de fabrication, des plus matures aux plus avancés. Si l'ASIC ne requiert pas de très hautes performances alors un procédé mature sera privilégié pour réduire les coûts. En revanche si le cahier des charges est très contraint alors un procédé très avancé sera privilégié. Un ASIC peut intégrer des parties numériques comme un FPGA mais également des parties analogiques, à base de composants élémentaires tels que des condensateurs, des résistances, des inductances, permettant ainsi de réaliser des fonctions mixtes. L'ASIC peut être optimisé également en consommation, car les procédés de fabrication intègrent aujourd'hui plusieurs types de transistors.

On peut donc utiliser des transistors très rapides pour les parties numériques devant effectuer des calculs en un temps le plus court possible, et utiliser des transistors faible consommation pour les parties ne nécessitant pas de rapidité d'exécution. Le choix d'un transistor est donc lié à un compromis entre vitesse et consommation, qui sont les 2 éléments clef dans la conception d'ASIC des systèmes actuels. Enfin, les circuits ayant pour vocation d'être fabriqués et utilisés en très grand nombre pour des applications grand public et/ou grand volume, sont des ASIC, car le coût de fabrication reste le plus avantageux. Ceci malgré la forte évolution des coûts liée à la complexité des procédés très avancés, type 28nm et en dessous par exemple.

Parmi les ASIC, on distingue 2 grandes catégories de circuits. D'une part les circuits purement numériques, pour lesquels on s'intéresse aux niveaux logiques '0' ou '1' des signaux, comme illustré par la figure 2.4, et pour lesquels les principales données du cahier des charges sont la vitesse et la consommation. La conception consiste à décrire le comportement du circuit dans un langage informatique puis à le convertir en masque de fabrication à l'aide d'outil de conception CAO très puissants et très sophistiqués.

FIG. 2.4 – *Signal numérique*

D'autre part les circuits dit analogiques pour lesquels on s'intéresse à l'aspect dynamique des signaux. Toutes les parties sont dimensionnées très finement une à une. Le cahier des charges peut intégrer des notions de vitesse et de consommation également, mais aussi de gain, de bande passante, de dynamique d'entrée ou de sortie, de résolution entre autres. Dans ce cas, les niveaux de tension peuvent prendre toutes les valeurs entre 0V et la tension d'alimentation, comme illustré par la figure 2.5.

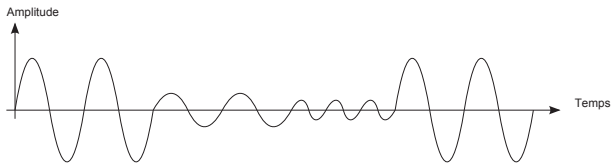


FIG. 2.5 – *Signal analogique*

Quelle que soit la nature du circuit, l'objectif est de réaliser la ou les fonctions en utilisant un minimum de surface pour évidemment réduire le coût unitaire. Dans la suite de ce manuscrit, nous vous présentons les 2 flots de conception pour chacune des catégories des ASIC, full custom et numérique, car l'approche, les outils et les

méthodes sont sensiblement différents. Chacun de ces flots utilise des fichiers technologiques inclus dans un kit de conception fourni par le fondeur. Ces notions seront indispensables pour comprendre la suite de ces travaux de thèse.

2.2 Flot de conception de circuit Full Custom

2.2.1 Conception et simulation électrique

La conception full custom consiste à prendre en compte un maximum de paramètres et d'éléments physiques. Il est important que le comportement de tous les composants, les tensions de tous les noeuds et les courants de toutes les branches soient pris en compte pour déterminer le comportement le plus proche possible de ce qu'il sera après fabrication sur silicium. La première étape est donc la conception du schéma de très bas niveau, c'est à dire au niveau transistors, intégrant d'une part un ou plusieurs type de transistors, MOS (Metal Oxide Semiconductor) ou bipolaire, mais également des composants spécifiques tels que les capacités, résistances, inductances entre autres. La figure 2.6 illustre un tel schéma d'une cellule composée de plusieurs composants élémentaires.

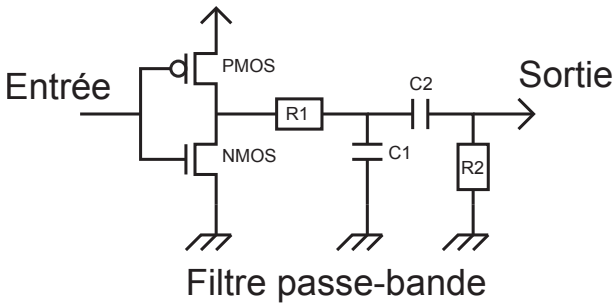


FIG. 2.6 – *Schéma d'une cellule full custom à base de composants élémentaires*

La conception d'une cellule permettant de réaliser une fonction logique à partir de transistors suit un mécanisme et des règles précises pour les procédés microélectroniques classiques qui sont des procédés CMOS, pour Complementary MOS. Cela consiste à définir la fonction elle-même à partir de transistors NMOS, c'est à dire les mettre en série pour une fonction "ET" logique et en parallèle pour une fonction "OU" logique, puis de faire l'inverse avec les transistors PMOS. Par exemple, pour implémenter la fonction logique "a.b+c", les 2 transistors NMOS commandés par les signaux 'a' et 'b' doivent être mis en série, eux-mêmes en parallèle avec le transistor NMOS commandé par le signal 'c'. Pour les transistors PMOS, le principe opposé est appliqué entre série et parallèle pour compléter le schématique de la fonction logique.

En technologie CMOS, la sortie est complémentée. Pour obtenir la fonction logique définie précédemment il est donc nécessaire d'ajouter un inverseur. Le schéma à base de transistors assurant la fonction logique "a.b+c" est présenté sur la figure 2.7

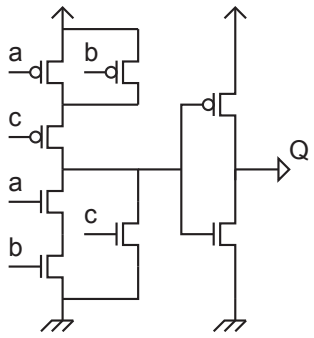


FIG. 2.7 – *Schéma d'une porte logique "a.b+c"*

Pour des circuits complexes, chaque bloc conçu peut être réutilisé plusieurs fois dans un autre schéma, couplés et connectés à d'autres blocs eux-mêmes conçus au niveau transistors. On parle alors de hiérarchie du circuit et d'instances. La profondeur de la hiérarchie est très variable en fonction de la complexité du circuit, mais également en fonction du besoin. Par exemple, dans un convertisseur analogique numérique flash, il est nécessaire d'instancier autant de comparateurs, lui-même conçu au niveau transistors, que de bits de résolution. La figure 2.8 illustre cet exemple, pour un CAN flash 7 bits, composé de 7 comparateurs et d'un encodeur lui-même étant un bloc numérique.

L'étape suivante consiste à simuler électriquement le comportement de la fonction réalisée. Il existe une multitude de simulations possible. Elle peut être fréquentielle pour observer par exemple des diagrammes de bode pour le gain et la phase, elle peut être statique pour analyser l'état dans lequel se trouvent les composants à l'équilibre, transistors MOS en régime bloqué, ohmique ou saturé par exemple. Il est également possible de faire des simulations transitoires, c'est à dire que l'on observe le comportement des tensions et des courants aux noeuds choisis ainsi que le comportement des signaux à tout moment. Dans certains cas, lorsque le cahier des charges impose des contraintes fortes, il est nécessaire de faire en plus des simulations de bruit. Toutes ces simulations sont basées sur 2 aspects. D'une part la modélisation des composants. Chacun d'entre eux est associé à un ou à plusieurs fichiers de description, communément appelé "modèle de simulation". Il tient compte de l'environnement dans lequel se situe le composant par rapport aux composants voisins, de la température, des tensions à ses bornes et des courants les traversant. Cette description est de façon générale basée sur des équations faisant appel à la physique. D'autre part, les simu-

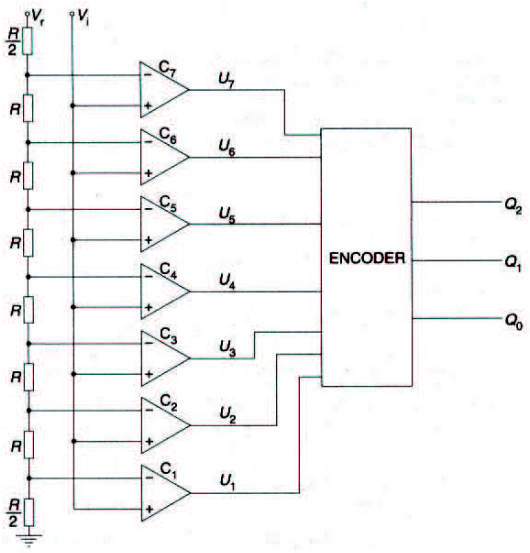


FIG. 2.8 – *Cellule à plusieurs niveaux de hiérarchie*

lations sont basées sur les outils logiciels et les simulateurs électriques notamment. Chacun d'entre eux ont des capacités à interpréter le contenu des modèles et à retranscrire le comportement. Le résultat d'une même simulation peut être différent selon le simulateur choisi. Généralement, les fabricants de circuits intégrés, appelés fonderies ou fondeurs, fournissent les modèles de simulation et recommandent également le simulateur à utiliser, car ils ont été préalablement validés. De plus, le langage de description et/ou la syntaxe ne sont pas les mêmes d'un simulateur à un autre. La figure 2.9 montre un résultat de simulation électrique, où l'on remarque bien le comportement dynamique des signaux.

A ce niveau-là de la conception il est également important de vérifier la robustesse d'une architecture. Une des raisons pour laquelle un circuit peut ne pas fonctionner, si les contraintes du cahier des charges sont très fortes par exemple, est la variation de taille pendant le procédé de fabrication. En effet, lorsque l'on dessine un transistor avec une largeur de 65nm par exemple, il se peut que sa taille réelle après fabrication soit différente de quelques pourcents. Chaque fondeur donne pour chaque niveau physique une fourchette dans laquelle il garantit la variation. Afin de prévoir ces variations de procédé de fabrication dès la conception, il est possible de faire des simulations dites de "Monte Carlo". Le principe est de lancer plusieurs simulations simultanément en faisant varier plusieurs paramètres de fabrication, soit par excès, soit par défaut. Le simulateur gère lui-même la variation de chacun des paramètres. Le concepteur choisit le nombre d'itérations, 500 par exemple, ainsi que l'amplitude de variation des paramètres. Souvent, plusieurs configurations de variation sont pro-

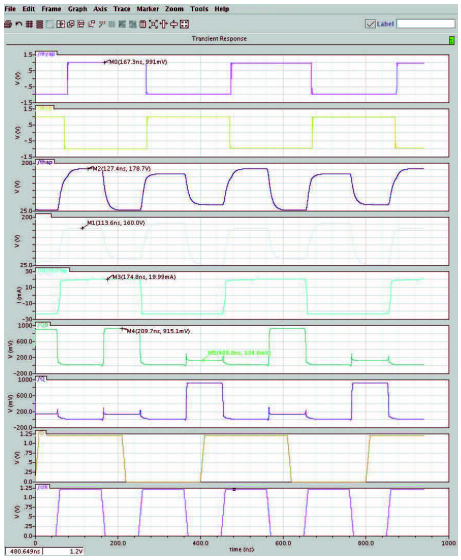


FIG. 2.9 – *Simulation électrique*

posées par les fichiers technologiques fournis par les fondeurs. Par exemple, SF pour Slow Fast ou FS pour Fast Slow. Cela correspond à une variation de +/- 3σ sur les Vtn et Vtp des transistors, soit à des transistors NMOS plus rapides et des PMOS plus lents, ou inversement. Le choix peut être SS pour Slow Slow ce qui correspond à +2σ sur les Vtn et -2σ sur les Vtp, ou encore FF pour Fast Fast, -2σ sur les Vtn et +2σ sur les Vtp. Ainsi le concepteur peut faire plusieurs types de simulation Monte Carlo pour s'assurer de la robustesse du circuit dans différentes configurations de variation de procédé de fabrication. Dans le cas de simulations transitoires, le résultat de cette étape se présente soit sous forme d'histogramme pour un instant t, soit sous forme de courbes toutes superposées les unes sur les autres.

2.2.2 Dessin des masques de fabrication

La fabrication d'un circuit microélectronique consiste à déposer plusieurs niveaux de couches conductrices et isolantes les unes sur les autres, sur une tranche de silicium appelé wafer, en plus des niveaux de dopage du wafer lui-même. Ce dernier subit préalablement plusieurs étapes de dopage local afin de définir principalement le type de transistors ainsi que leurs sources et drains. Pour la fabrication, chacun de ces niveaux nécessitent d'avoir un ou plusieurs masques, typiquement de quartz. Afin de pouvoir fabriquer ces masques, ils doivent être dessinés par le concepteur du circuit. Le dessin des masques d'un circuit complet est communément appelé "layout", qui est au final un empilement et une juxtaposition de polygones. La figure 2.10 présente le dessin des masques d'un inverseur simple et celui d'une NAND à 2 entrées. On

remarque que le layout de la NAND est relativement complexe, car il s'agit d'une cellule rapide donc avec des buffers de sortie importants.

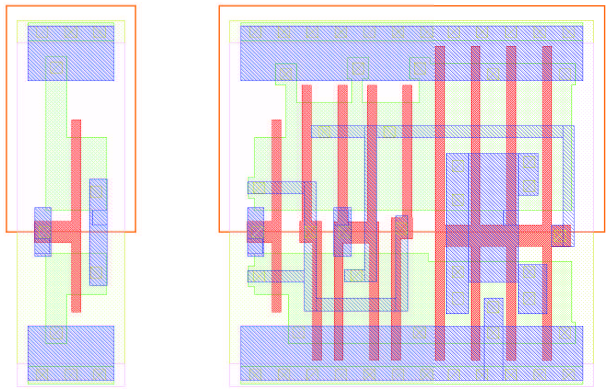


FIG. 2.10 – *Dessin des masques d'un inverseur et d'une NAND_X8*

Pour concevoir le dessin des masques d'un bloc full custom, soit chacun des polygones est traditionnellement dessiné individuellement, ce qui n'est presque plus vrai aujourd'hui. Effectivement il est très pratique voire indispensable, d'avoir accès à des cellules paramétrables pour chacun des composants élémentaires. En effet, ce type de cellule permet de dessiner automatiquement le dessin de tous les masques constituant les transistors, résistances et autres éléments. L'étape suivante étant de placer judicieusement tous ces composants et de les interconnecter, manuellement. Ces cellules paramétrables offrent également l'avantage de pouvoir changer la taille d'un composant dynamiquement, tout en respectant les règles de dessin de la technologie pour laquelle le fabricant fournit la cellule paramétrable.

De plus, les suites d'outils logiciels industriels permettent de générer le dessin des masques de tous les composants placés sur la vue "schematic" vers la vue "layout" de façon automatique. Il est ensuite possible de faire apparaître un chevelu sur la vue layout entre les différents terminaux des composants, c'est à dire d'afficher de façon symbolique toutes les connexions devant être réalisées entre les composants pour être conforme au schéma. Chaque fois qu'une connexion est faite entre 2 terminaux, le fil correspondant du chevelu disparait, si bien qu'à la fin de la conception du layout plus aucun fil du chevelu n'apparait.

2.2.3 Vérification DRC

Pour fabriquer des circuits intégrés, les machines et équipements nécessaires sont plus ou moins sophistiqués, selon les fonderies, les étapes à réaliser, le noeud technologique. Dans tous les cas, les machines ont des limites physiques à respecter pour

assurer une bonne fabrication des wafers et un fonctionnement des circuits conforme aux attentes, bien qu'il y ait une part importante due à la robustesse de la conception. Ces règles sont définies et imposées par les fonderies, dans un manuel appelé DRM, pour Design Rule Manual. On y retrouve toutes les règles avec d'une part un nom de code unique pour chacune d'entre elle, des valeurs minimums, maximums ou fixes, ainsi qu'une illustration de l'erreur sur un exemple de dessin de masque.

Les erreurs peuvent être entre 2 polygones du même masque dans le cas d'un espacement par exemple, ou entre 2 polygones de masques différents dans le cas d'un composant. Elles peuvent être locales, comme les règles de taille minimum voire encore d'encombrement, mais elles peuvent aussi être globales, comme les règles de densité sur les différents niveaux de métaux. Quelques règles, les plus courantes, sont illustrées sur la figure 2.11.

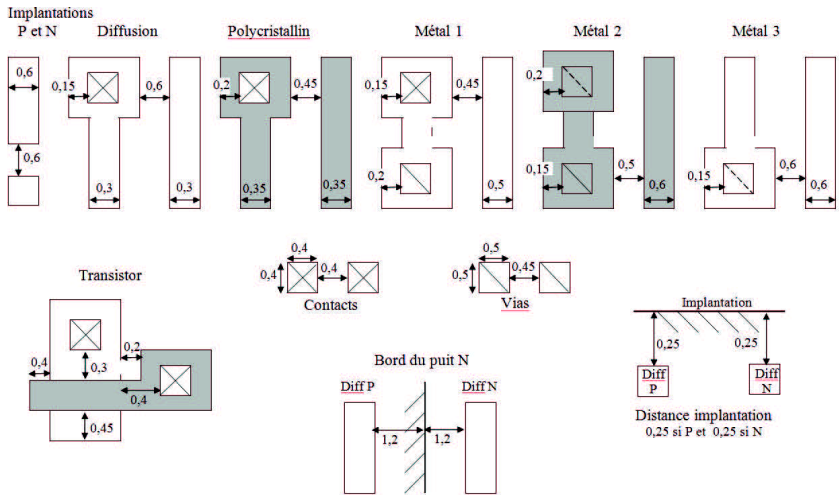


FIG. 2.11 – *Règles DRC les plus courantes*

La vérification de ces règles se fait à partir des outils de conception, de façon automatisée. Chaque outil possède son propre langage pour décrire ces règles qui sont toutes implémentées dans un fichier technologique. Généralement, les fonderies fournissent les fichiers technologiques pour plusieurs outils de conception, au moins pour les plus utilisés dans le monde de l'industrie, ou alors en fonction des besoins des utilisateurs.

Le tableau 1.1 dont les données ont été extraites à partir des technologies des fondeurs ams et STMicroelectronics, montre qu'à procédé de gravure équivalent le nombre de règles à vérifier est plus important pour un procédé bipolaire que pour un procédé CMOS. Ceci est directement lié au nombre de masques, car en effet, un procédé bipolaire compte plus de masques qu'un procédé CMOS. Il montre également

l'évolution du nombre de règles en fonction de la finesse de gravure. Le nombre de règles est très dépendant de la complexité du procédé de fabrication. Plus il est avancé, plus le nombre de masques augmente. Il en est de même pour le nombre de règles DRC.

CMOS	**0.35u**	**130n**	**90n**	**65n**	**45n**
	400	750	830	1530	2620
BiCMOS	**0.8u**	**0.35u SiGe**	**0.35u 6G**	**0.25u**	**0.13u**
	240	495	665	850	1660

TAB. 2.1 – *Nombres de règles DRC vérifiées par procédé de fabrication*

Dans les années 1990, lorsque les technologies les plus avancées offraient une taille minimum de largeur de transistors de 0.8u, le nombre de règles à vérifier étaient d'environ 200 pour un procédé CMOS et environ 250 pour un procédé BiCMOS. De nos jours, une dizaine de générations plus tard, les technologies les plus avancées offrent la possibilité de dessiner des grilles de transistors de 28nm, voire 20nm. Le nombre de règles à vérifier est considérable, de l'ordre de 3 000 règles. Une autre raison qui explique cette croissance faramineuse est que la taille des wafers, qui était de 100 mm dans les années 1990, est aujourd'hui de 300 mm, ce qui impose certaines contraintes sur la conception des circuits. Quelques fondeurs commencent même à travailler sur des wafers 450 mm.

2.2.4 Vérification LVS

Le schéma a pour but de décrire les interconnexions entre les composants pour valider le fonctionnement d'un circuit intégré par simulation électrique et de s'assurer que le cahier des charges est bien respecté du point de vue des performances souhaitées. Le layout quant à lui a pour but de fournir la base de données nécessaire à la fabrication des circuits intégrés sur les tranches de silicium. Il est donc tout à fait indispensable de faire le lien entre les 2 et de s'assurer que le dessin des masques correspond bien au schéma qui a servi aux simulations, bien que cette tâche puisse être facilitée par l'utilisation du chevelu. Pour cela, il est nécessaire de faire une vérification LVS: Layout Versus Schematic. Le principe est le suivant: à partir du schéma, l'outil de conception est capable d'écrire une "netlist", qui est un fichier listant tous les composants de base utilisés ainsi que toutes les interconnexions entre eux. Pour cela, il s'appuie sur d'une part tous les paramètres définis dans les cellules paramétrables par exemple, et d'autre part sur les "cdf" des composants. Le "cdf", pour Component Description Form" défini pour chaque vue disponible et pour chaque

composant, les éléments à extraire. Par exemple, un transistor NMOS possède une vue auLVS, pour laquelle il est défini dans le "cdf" d'extraire les paramètres suivants: la longueur, la largeur, la surface et le périmètre. Dans ce fichier, la hiérarchie peut être conservée, ce qui permet d'identifier chacun des blocs du circuit. Du point de vue dessin des masques, la tâche est nettement moins évidente, car l'outil doit extraire le même type de fichier que pour le schéma, mais à partir des masques et des polygones définis sur le layout par le concepteur. On parle alors dans le flot de conception de "l'extraction". Cette étape nécessite d'avoir un fichier technologique spécifique, décrit dans un langage propre à l'outil logiciel de conception, fourni à nouveau par le fondeur. Dans ce fichier technologique, une section est dédiée à la définition de tous les niveaux disponibles sur le procédé de fabrication, communément appelé "layers". Une autre section est dédiée à la définition de layers dérivés, basés sur les layers de base et sur des commandes spécifiques du langage. Par exemple, un transistor étant formé entre le croisement d'un polygone de silicium poly cristallin et un polygone de zone active de diffusion, les commandes correspondantes sont:

POLY = *geomGetPurpose("poly" "drawing")*
ACTIVE = *geomGetPurpose("active" "drawing")*
NPLUS = *geomGetPurpose("nplus" "drawing")*
PPLUS = *geomGetPurpose("plus" "drawing")*

Ces 4 layers *POLY, ACTIVE, NPLUS et PPLUS* sont des niveaux de base correspondant à des masques de fabrication.

NMOS = *geomInside(geomAnd(POLY ACTIVE) NPLUS)*
PMOS = *geomInside(geomAnd(POLY ACTIVE) PPLUS)*

Ces 2 layers *NMOS et PMOS* sont des niveaux dérivés n'étant utilisés que du point de vue logiciel pour le LVS.

Dans ce fichier technologique utilisé pour l'extraction, une autre section décrit tous les composants disponibles dans les bibliothèques fournies par le fondeur. Chacun des composants fait appel à plusieurs layers dérivés, car il est nécessaire d'une part de définir chaque composant, mais également tous leurs terminaux et de façon individuelle. Ceci permet d'avoir une cohérence dans l'extraction de la "netlist", et de pouvoir déterminer à quel potentiel est connecté chacun d'entre eux. En effet, si un ou plusieurs terminaux sont mal définis, alors la liste des noeuds pourrait être fausse et par conséquent le LVS indiquerait des erreurs alors qu'il n'y en aurait pas. Enfin,

cette étape permet d'extraire les paramètres des composants, comme la longueur et la largeur des transistors et des résistances. Ces paramètres seront également comparés à ceux du fichier extrait du schématique. Si les 2 "netlists" issues du schématique et du dessin des masques sont conformes, on dit que le LVS "match".

2.2.5 Simulation post-layout

L'extraction permet d'une part d'extraire la liste des noeuds et des composants en vue du LVS, comme nous l'avons décrit dans le paragraphe précédent, mais elle permet également d'extraire les composants parasites depuis le dessin des masques. Le schéma étant basé sur des symboles et des connexions abstraites, le layout au contraire fait appel à des grandeurs physiques. Il comporte bien entendu tous les composants mais également toutes les connexions qui seront réalisées et fabriquées, avec leur taille finale. Chaque piste métallique d'interconnexion entraine d'une part une résistance parasite série entre les 2 terminaux qu'elle connecte, et d'autre part une ou plusieurs capacités parasites, soit avec une piste au-dessus de celle-ci, soit avec une piste au-dessous, soit avec le substrat, soit avec une piste à côté. La figure 2.12 donne un aperçu de toutes les capacités parasites que l'on peut retrouver sur un procédé de fabrication à 6 niveaux de métal.

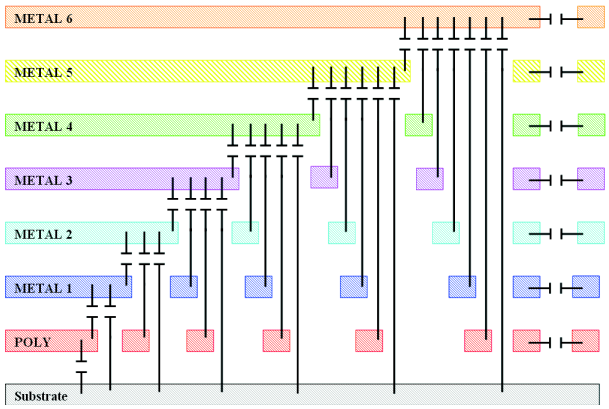


FIG. 2.12 – *Capacités parasites inter-métal*

Il faut éventuellement ajouter à cela toutes les capacités de sources, grilles et drains qui sont la plupart du temps déjà modélisées dans les modèles de chaque composant. En conséquence, celles-ci n'apparaissent pas dans le fichier d'extraction.

Le fait d'avoir la possibilité d'extraire les capacités parasites d'un circuit est très important, surtout lorsque les contraintes fixées dans le cahier des charges sont fortes, car ces composants non souhaités peuvent altérer gravement le fonctionnement du circuit. Les problèmes les plus fréquents sont ceux liés aux retards induits sur les

signaux. En effet, lorsqu'un signal doit charger une multitude de capacités parasites sur un noeud, bien que très petites, les temps de propagation s'allongent. L'effet des résistances augmente ce phénomène. Dans des circuits analogiques comportant des tensions de référence ou de polarisation très précises par exemple, le comportement peut aussi être altéré par ces composants parasites. C'est pourquoi il est courant de faire des simulations non pas seulement du schéma, mais également de la "netlist" comportant tous les parasites.

De la même façon que pour les composants de base, les outils de conception s'appuient sur des fichiers technologiques permettant de reconnaître chacun de ces composants, et d'en extraire leur valeur. Les valeurs typiques sont de l'ordre de quelques atto Farad ($1aF = 1 \times 10^{-18}\ F$) à quelques femto Farad ($1fF = 1 \times 10^{-15}\ F$). Le calcul de ces capacités et directement lié à l'épaisseur des différents métaux et à l'épaisseur des diélectriques qui les séparent.

2.3 Flot de conception de circuits numériques

Comme précisé précédemment, un circuit purement numérique utilise des signaux logiques, soit '1' pour Vdd, soit '0' pour Gnd, basés sur des équations booléennes utilisant des portes logiques. On parle alors de signaux binaires. Une application typique d'un circuit numérique pourrait être le processeur d'un ordinateur. Ce type d'application est souvent caractérisé par une fréquence, au moins sur le marché grand public où l'on voit la fréquence des processeurs augmenter sans cesse. Mais dans la réalité, au niveau de la conception, les contraintes ne sont pas seulement celles-ci. Un circuit numérique performant doit effectivement être rapide, mais il doit être peu gourmand en énergie. A l'heure où nous utilisons dans la société moderne de plus en plus d'applications embarquées et/ou autonome tels que les PDA ou les téléphones portables, ces circuits se doivent de consommer un minimum de puissance.

En termes de conception, l'approche est tout à fait différente de celle des circuits analogiques ou plus généralement des circuits "full custom". Nous allons à travers la deuxième partie de ce chapitre, décrire l'ensemble du flot de conception d'un circuit numérique, ce qui permettra de bien appréhender la suite des travaux effectués à travers cette thèse sur le développement d'un kit de conception et flot de conception pour technologie hybride CMOS/Magnétique.

2.3.1 Description comportementale

La conception d'un circuit numérique consiste à décrire son fonctionnement par un langage de programmation. On ne parle donc ni de transistor ni de composant

à manipuler, au moins jusqu'à l'étape de conception du dessin des masques. Deux langages sont principalement utilisés: le VHDL, majoritairement utilisé en Europe, et le Verilog, souvent préféré de l'autre côté de l'atlantique. Ces 2 langages, qui sont des langages de description de matériel, répondent à des besoins similaires. Il existe beaucoup de programme de conversion de l'un vers l'autre. Ces langages permettent de garder une certaine hiérarchie dans le circuit et de construire une structure très précise. Dans un processeur par exemple, plusieurs modules seront décrits et implémentés séparément, tels que l'unité arithmétique et logique (UAL ou ALU), les décodeurs, les différentes machines à états et bien d'autres. Chacun de ces modules peut être décrit et simulé individuellement. Ils seront ensuite instanciés au plus haut niveau de la hiérarchie et connectés entre eux.

Ces langages, le VHDL en particulier, ont plusieurs objectifs, à plusieurs niveaux dans la conception. Il est possible par exemple d'intégrer dans un premier temps des notions de timing entre les signaux lorsque l'objectif est la simulation fonctionnelle dite globale de plus haut niveau. A cette étape, le comportement du circuit est tout à fait symbolique et ne prend pas en compte la technologie, c'est à dire les performances et contraintes du procédé de fabrication. En revanche, lorsque l'objectif est de réaliser un circuit en vue de la fabrication, il n'est plus possible d'intégrer ces notions de timing entre les signaux eux-mêmes. Il existe un certain nombre de règles à respecter qui consistent à décrire le circuit en vue de la synthèse logique (cf. section "Synthèse logique"). On dit alors que le VHDL est synthétisable ou du type RTL: Register Transfer Level. Un concepteur très expérimenté sait exactement qu'elle est l'influence du code qu'il écrit sur la façon dont le circuit sera implémenté par les outils en termes de matériel, d'architecture et de porte logique.

Notons tout de même que cette phase de conception est la même pour un ASIC que pour un FPGA, dans la mesure où cette étape n'est liée ni à une technologie ni à un mode d'implémentation.

2.3.2 Simulation comportementale

La première phase de la conception d'un circuit numérique étant la description comportementale, il paraît évident que l'étape suivante consiste à vérifier son fonctionnement, par simulation logique. Dans ce cas-là, aucun aspect de timing n'est pris en compte, seul l'état des niveaux des sorties est analysé, ainsi que la cohérence entre les différents modules. A l'issu de cette étape, le concepteur est capable de dire si le circuit décrit en VHDL ou en Verilog est conforme au fonctionnement souhaité, mais il n'est pas encore capable de dire si les contraintes imposées par le cahier des charges sont respectées. La figure 2.13 montre un résultat de simulation logique où

l'on voit toutes les transitions de chaque signal où tous les signaux sont binaires.

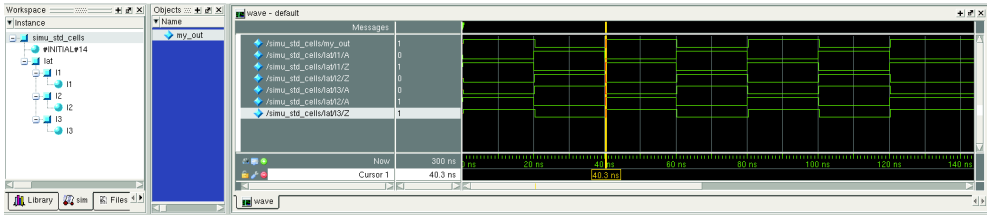


FIG. 2.13 – *Simulation numérique*

2.3.3 Synthèse logique

A partir de cette étape, le concepteur doit faire son choix entre non seulement concevoir un ASIC ou un FPGA, mais surtout la technologie de fabrication. En ce qui concerne l'option ASIC, le procédé de fabrication est un procédé CMOS. Le choix se porte alors sur la performance pouvant être atteinte. Il y a un compromis à trouver entre vitesse, consommation et coût. Généralement, plus le procédé de fabrication est avancé, 28n ou 40n par exemple, plus les vitesses atteignables sont importantes, mais plus les consommations dynamique et statique sont élevées également. En ce qui concerne le coût, les technologies très avancées sont plus chères que les plus matures, du fait que leur complexité qui nécessitent des équipements très performants. Néanmoins, elles permettent une densité d'intégration bien plus importante, ce qui compense la hausse du coût. Selon l'application et la configuration du circuit, ce coût peut être équivalent entre une technologie 28n ou 40n. Le tableau 2.2 donne un ordre de grandeur des densités d'intégration des procédés de fabrication CMOS des 15 dernières années.

On remarque sur ce tableau une évolution tout à fait impressionnante. En moins de 20 ans, les recherches et les efforts menés au niveau des procédés de fabrication et sur les équipements ont permis de réduire de presque 30 la taille minimum des transistors et de multiplier par 2 500 le nombre de portes logiques qu'il est possible de placer dans 1 mm^2 de silicium. Par ailleurs, un paramètre important dans le choix de la technologie est le nombre d'entrées et de sorties du circuit, car les plots de bonding occupent une part absolument pas négligeable dans un circuit intégré.

Lorsque ce choix de la technologie est fait, le concepteur peut alors faire la synthèse logique. Cette étape est en fait décomposée en deux parties. La première partie consiste à charger les fichiers VHDL dans l'outil de synthèse qui va alors vérifier la structure du point de vue du langage, et pour les circuits complexes ayant une hiérarchie, l'outil vérifie si tous les blocs sont présents. Une première synthèse est alors

CMOS 0.8 μm	=	1.2K	portes / mm^2
CMOS 0.6 μm	=	3K	portes / mm^2
CMOS 0.35 μm	=	18K	portes / mm^2
CMOS 0.25 μm	=	35K	portes / mm^2
CMOS 0.18 μm	=	80K	portes / mm^2
CMOS 0.13 μm	=	180K	portes / mm^2
CMOS 90nm	=	400K	portes / mm^2
CMOS 65nm	=	800K	portes / mm^2
CMOS 40nm	=	1600K	portes / mm^2
CMOS 28nm	=	3000K	portes / mm^2

TAB. 2.2 – *Densité d'intégration des procédés de fabrication CMOS*

faite. Elle consiste à convertir une description comportementale décrite en langage matériel en un ensemble de portes logiques interconnectées entre elles. Cette première synthèse est générique, c'est à dire que les cellules utilisées par l'outil ne sont pas encore celles des bibliothèques du procédé choisi par le concepteur, mais des cellules de bibliothèques génériques. Seule la fonction logique des cellules est utilisée pour transcrire la fonction globale du circuit. A l'issue de cette première partie, on peut distinguer chaque bloc de la hiérarchie. On a accès à son symbole, ce qui permet de voir tous les terminaux d'entrée et de sortie, ainsi qu'à son schéma. Le schéma n'a donc qu'un seul type de porte NAND, un seul type d'inverseur etc. La figure 2.14 montre un exemple de synthèse basée sur des composants génériques.

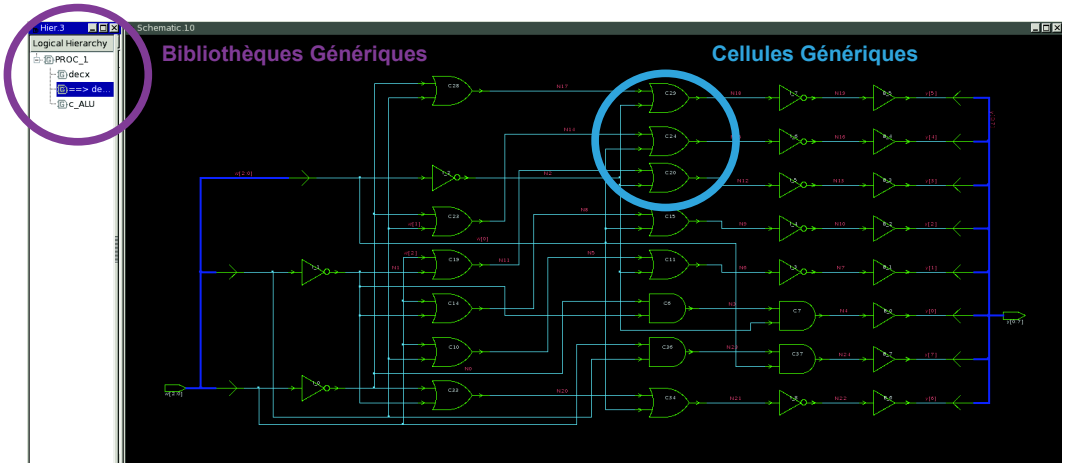


FIG. 2.14 – *Synthèse logique générique*

C'est seulement lors de la deuxième partie de la synthèse que l'outil va remplacer toutes les cellules génériques par des cellules de bibliothèques de la technologie choisie, appelées cellules standards, fournies par la fonderie. L'outil analyse principalement 2 aspects: tout d'abord que toutes les cellules génériques puissent être remplacées simplement par des cellules standards. Par exemple, si lors de la première synthèse l'outil a utilisé une cellule AND et qu'aucune AND n'est disponible dans les bibliothèques standards, alors il va refaire un calcul de décomposition d'équation logique pour utiliser une cellule NAND associée à une cellule "inverseur". De même, si l'outil a utilisé une bascule avec set, reset et enable et que ces éléments ne sont pas disponibles dans le kit de conception, l'outil est capable de refaire un calcul sur les équations booléennes pour utiliser les cellules disponibles. En revanche, les bibliothèques étant souvent très fournies et contenant des cellules complexes du type AND-OR-INV, l'outil peut simplifier le schéma et n'utiliser qu'une seule cellule complexe pour remplacer plusieurs cellules génériques. La figure 2.15 donne un exemple de synthèse basée sur des cellules standards de bibliothèque de fonderie. On dit que la synthèse est "mappée" sur une bibliothèque standard.

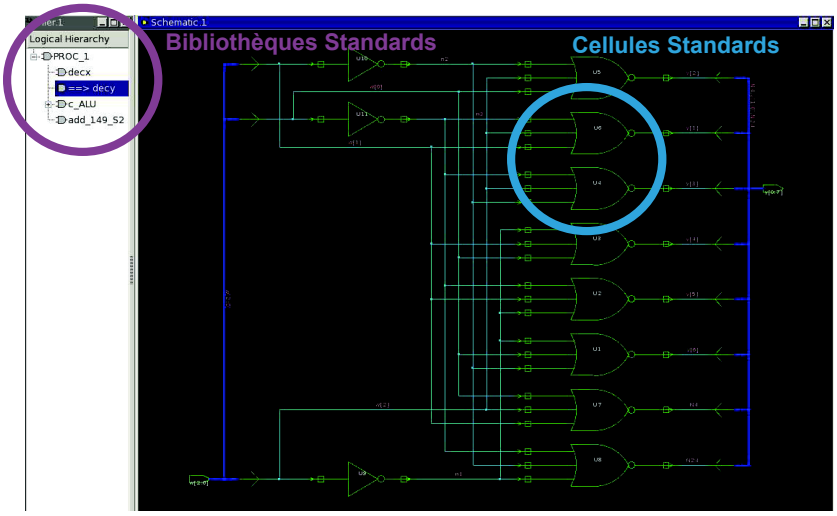


FIG. 2.15 – *Synthèse logique mappée sur une bibliothèque standard*

Le deuxième aspect analysé par l'outil est la taille des cellules à utiliser parmi toutes celles de disponible. Par exemple, s'il doit remplacer un inverseur et qu'il en a plusieurs de disponibles, il doit faire un choix. Le choix se fait en fonction de la cellule ou des cellules connectées à sa sortie. Prenons l'exemple de la figure 2.16. Plus il y a de cellules connectées en sortie de la bascule, plus la charge de cette bascule dues aux capacités d'entrées des NAND, NOR et INV sera élevée. La sortie de la bascule doit donc être dimensionnée de façon à ce que les temps de propagation du signal

"sig" permettent de respecter les contraintes de temps imposées par le concepteur, dans la mesure du possible.

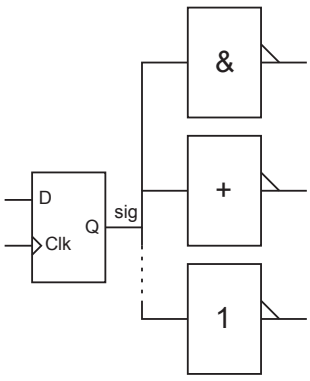


FIG. 2.16 – *Dimensionnement de la sortance d'une porte logique*

Généralement, les bibliothèques standards proposent plusieurs versions de chaque cellule, avec différentes sortances. Plus les bibliothèques sont riches en cellules, plus le résultat de la synthèse sera optimisé, en surface, timing et consommation. Pour homogénéiser les temps de propagation et pour respecter les contraintes sur la fréquence de l'horloge, dans le cas de circuit synchrone, la cellule avec la sortance adéquate doit être utilisée. La cellule ne doit ni être sous dimensionnée pour ne pas pénaliser les performances du point de vue de la vitesse, ni être sur dimensionnée pour n'engendrer ni de surconsommation inutile ni une augmentation de la surface. C'est pourquoi il est important de bien définir les contraintes dans l'outil de synthèse pour réellement optimiser le circuit en fonction des besoins et contraintes. De plus, il est possible de faire une nouvelle synthèse avec insertion de buffer, après placement sur le layout, afin d'optimiser les aspects de timing et de sortance, réduire les chemins critiques et améliorer l'efficacité de l'arbre d'horloge.

Si pour une raison ou une autre il n'y avait pas assez de cellules disponibles et que la synthèse ne pouvait pas aboutir, alors le ou les blocs n'ayant pas été synthétisés apparaissent toujours de la même manière qu'à la fin de la première synthèse, c'est à dire avec un "G" signifiant que les cellules sont toujours issues de la bibliothèque Générique. C'est ce que l'on peut remarquer sur la figure 2.14.

Lorsque la synthèse est terminée, l'outil logiciel est capable d'extraire plusieurs informations de performances concernant le circuit, certaines sous forme textuelle, d'autres sous forme graphique. On peut connaitre non seulement le nombre de cellules utilisées et la surface totales de toutes ces cellules, mais également des informations sur les consommations dynamique et statique, comme illustré sur la figure 2.17.

De plus, il est possible d'extraire des données sur les temps de propagation et

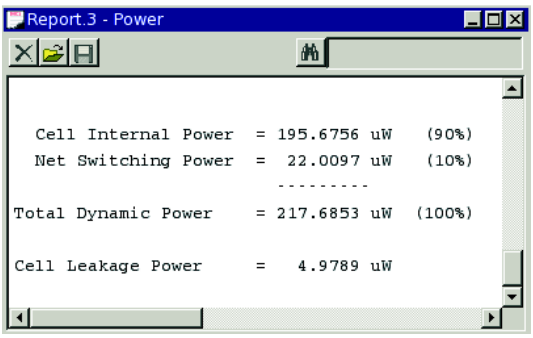


FIG. 2.17 – *Rapport de consommation après synthèse*

ainsi de connaitre quelle sera la fréquence d'horloge maximum atteignable pour le circuit. Le concepteur peut alors analyser quel sera le chemin critique, c'est à dire le chemin pour lequel le temps de propagation et de stabilisation d'un signal prendra le plus de temps, entre l'entrée et la sortie d'une partie combinatoire. La figure 2.18 montre le temps de propagation de tous les signaux du circuit. Ceci est appelé le "path slack", c'est à dire le temps restant entre la fréquence d'horloge souhaitée et indiquée à l'outil avant la synthèse en contrainte, et le temps de propagation réellement nécessaire après la synthèse.

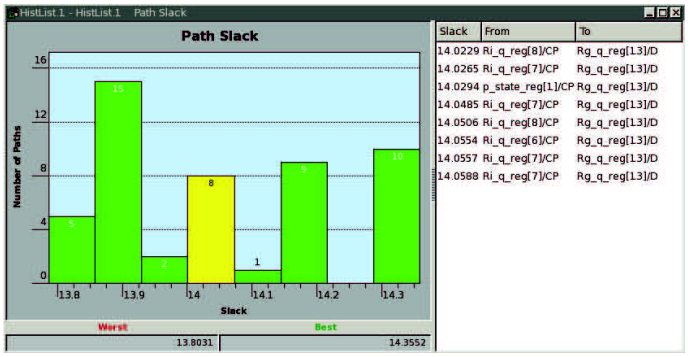


FIG. 2.18 – *Rapport de synthèse sur le timing: path slack*

Si le slack est positif alors il n'y a pas de problème de timing. S'il est négatif oui. Pour que le circuit soit optimisé en fréquence l'idéal est que le slack soit homogène, ou au moins qu'il n'y ait pas de grand pic pour 1 seul ou 2 signaux. Si tel est le cas, le concepteur peut alors refaire plusieurs synthèses pour améliorer toutes ces performances, mais il peut également modifier le code source VHDL de certains modules, en découpant d'avantage les blocs et en ajoutant des étages de "pipeline" pour diminuer un chemin critique qui serait très long par rapport aux autres temps de propagation. Après optimisation du slack, la fréquence maximum de fonctionnement

est alors déterminée.

Pour être capable de faire toutes ces opérations, de mapper une synthèse de portes logiques génériques sur une technologie choisie, calculer la sortance nécessaire en fonction des connexions sur les noeuds de sortie, extraire des informations sur la performance du circuit synthétisé, l'outil de synthèse s'appuie à nouveau sur des fichiers technologiques, de bibliothèques cette fois-ci. Ils sont communément appelés ".lib". Dans ce type de fichier, sont déclarés tous les composants, tous leurs terminaux, la fonction de chacun des terminaux (clock, entrée, sortie). De plus, leur surface, leur consommation statique, leur sortance et la capacité interne de chaque terminal, entre autres, sont déclarés dans ce fichier technologique. Il est donc important que ces fichiers soient très bien définis pour avoir une synthèse optimisée et fiable. Ces données sont extraites par simulations électriques au niveau transistor, comme présenté dans la section "conception full custom", pour chacune des cellules standards disponibles. On parle alors de caractérisation de cellules de bibliothèques.

2.3.4 Simulation de circuit après synthèse

La première simulation dans le flot de conception d'un circuit numérique consiste à vérifier le bon fonctionnement du point de vue sémantique où seuls les niveaux logiques sont pris en compte. Mais il est indispensable de s'assurer du fonctionnement du circuit après synthèse, car cela reflète plus précisément le circuit qui va être fabriqué. Pour cela, l'outil de simulation utilise des fichiers technologiques qui décrivent le comportement exact de chaque cellule de bibliothèque. Cette description se fait au format Verilog pour lequel il existe 2 façons de décrire le comportement des cellules. Soit par équations booléennes, pour les cellules purement combinatoires, soit sous forme de table de vérité impliquant des changements d'états sur certains signaux, pour les cellules séquentielles, comme présenté ci-dessous.

```
posedge CP => (Q +: D)) =
        if  (!En) (posedge Clk => (Q +: D)) = (0.1, 0.1);
        if  (En)  (posedge Clk => (Q +: '0')) = (0.1, 0.1);
```

On voit apparaitre dans cette description des éléments de timing de 0.1 ns, mais qui ne sont que des valeurs arbitraires figées, et qui ne correspondent pas à la réalité. Tous les aspects liés aux temps de propagation sont définis dans le fichier ".lib" évoqué précédemment dans lequel se trouvent par exemple les temps de montée et de descente pour chaque terminaux, les règles sur les temps de maintien au niveau haut et niveau bas à respecter pour le bon fonctionnement de la cellule. Toutes ces données, très précises, et qui permettent d'avoir des rapports de synthèse précis, sont

déterminées au préalable par simulation électrique. En effet, chaque cellule standard est conçue dans un premier temps selon le flot de conception "full custom" présenté au début de ce chapitre. Toutes les valeurs extraites sont donc basées sur les résultats obtenus à partir des modèles de simulation décrivant le comportement de chaque transistor. Le principe est alors de générer un fichier SDF, Standard Delay Format, depuis la description du circuit après synthèse, ce qui permet d'extraire les paramètres de timing de chaque cellule. Ce fichier est ensuite injecté dans la "netlist" lors de la simulation, qui est alors réaliste du point de vue timing par rapport au circuit fabriqué. On dit que l'on fait une simulation avec rétro-annotation.

En ce qui concerne la conformité de cette "netlist", il est possible à l'issue de cette phase de simulation, de l'importer dans un éditeur de schéma. L'outil va créer une vue "schematic", comportant toutes les cellules utilisées, elles-mêmes ayant toutes une vue "schematic", ainsi que leurs connexions. Le but étant par exemple de faire une vérification LVS ou d'utiliser le bloc numérique dans un circuit mixte, qui peut être connecté à d'autres parties analogiques par exemple.

2.3.5 Dessin des masques: Placement et Routage

La phase de conception du dessin des masques est également très différente de celle des circuits full custom. En effet pour chaque cellule de bibliothèque sont fournis d'une part une vue "layout" qui comportent tous les masques nécessaires à la fabrication et d'autre part une vue "abstract". Contrairement à la vue layout, qui ne peut pas toujours être distribuée si elle est considérée par le fondeur comme propriété intellectuelle (IP) ne pouvant être divulguée, la vue abstract est beaucoup plus réduite. Elle ne comporte que les polygones du premier niveau de métal servant aux interconnexions, c'est à dire les pins d'entrée et de sortie, ainsi que les 2 rails d'alimentations Vdd et Gnd. La figure 2.19 montre la différence en ces 2 vues.

Pour l'étape de placement et de routage l'outil CAO utilise le fichier de description du circuit après synthèse, qui décrit tout le circuit à partir des cellules standards utilisées et toutes les interconnexions entre elles. Cette étape consiste à placer toutes ces cellules dans un espace alloué, puis à effectuer le routage de tous les signaux entre terminaux. Ceci se fait en suivant un flot relativement standardisé, dont voici les principales étapes à effectuer:

– Définir l'encombrement, appelé "floorplan". Cela consiste à imposer une taille et une géométrie au circuit, soit carrée soit rectangulaire. C'est également à cette étape que l'on peut choisir de placer certains modules de la hiérarchie stratégiquement les uns par rapport aux autres pour gagner en performance.

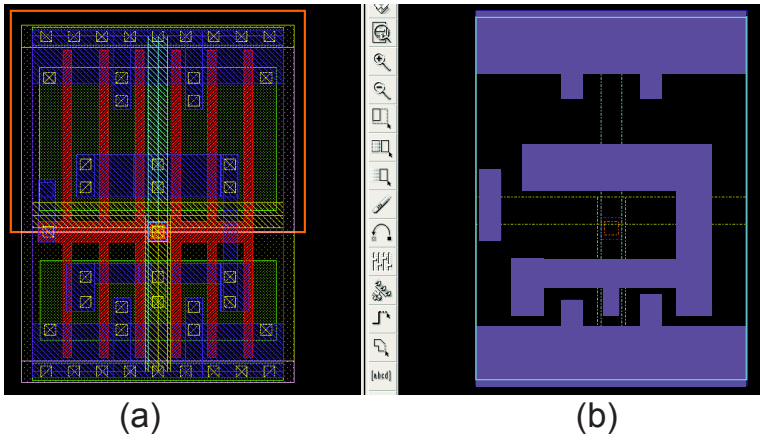


FIG. 2.19 – *Vues Layout (a) et Abstract (b) d'une cellule standard*

On définit aussi l'encombrement à l'extérieur du circuit pour les rails d'alimentation.

– Définir l'arbre d'alimentation pour les rails Vdd et Gnd, ainsi que leur taille. Chacun de ces rails est connecté à des plots d'alimentation. Ces connexions véhiculent tout le courant qui circulera dans le circuit, et doivent donc être dimensionnées en conséquence. On dit que pour une technologie pour laquelle les métaux sont principalement de l'aluminium, la taille doit être de 1 μm / 1mA DC, et de l'ordre de 1 μm / 10mA DC pour des pistes en cuivre. Ces rails sont distribués vers le coeur du circuit, horizontalement et potentiellement verticalement si besoin, pour multiplier les connexions aux cellules standards. Le but est d'éviter d'une part les problèmes liés à l'électro migration et d'autre part d'éviter les chutes de tension dues aux longueurs de fils trop importantes. Une chute de tension peut entrainer un dysfonctionnement car les temps de propagation se rallongent en conséquence, et donc le timing entre les blocs peut ne plus être respecté.

– Placer toutes les cellules à l'intérieur de chaque module si le circuit est complexe et/ou hiérarchique, ou les placer sur l'ensemble du floorplan alloué. Toutes les cellules sont donc alignées par rangées et retournées de 180° une rangée sur deux, pour une optimisation de surface. Cela implique que chaque cellule soit conçue avec les mêmes contraintes de hauteur et de taille de pistes pour les rails d'alimentation, afin de permettre ce type de placement automatique. Après cette étape, subsiste une multitude de plus ou moins petits espaces entre les

cellules standards, comme illustré sur la figure 2.20, qui doivent être comblés. Ceci est fait à l'aide de cellules dites "Filler". Le fondeur fournit plusieurs taille de Filler ce qui permet de combler tous les espaces vides et d'avoir une densité de 100% de la surface.

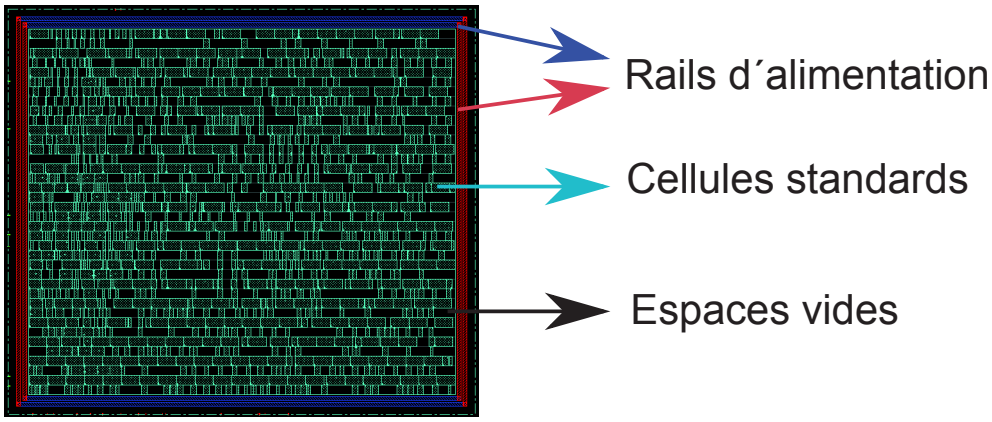


FIG. 2.20 – *Placement automatique des cellules standards*

– Router l'ensemble des signaux. Pour des circuits où les contraintes sont fortes en termes de fréquence d'horloge, il est courant de faire le routage en 2 étapes. Tout d'abord, router le signal d'horloge seulement. On parle communément "d'arbre d'horloge". Le principe est semblable à celui des pistes d'alimentation. Cette étape permet de router le signal d'horloge en évitant au maximum de changer de niveaux physiques dans les couches d'interconnexion, ce qui ajouterait une multitude de vias de connexion, donc des résistances série, ce qui augmenterait les temps de propagation. Ensuite, il suffit de router tous les autres signaux du circuit.

Notons tout de même qu'il est tout à fait possible d'inclure dans ce flot de placement et de routage les plots de bonding utilisés pour la mise en boîtier. L'outil sait placer judicieusement tous les plots, ajouter des cellules "Filler" si nécessaire dans la couronne de plots et les router vers le coeur du circuit.

Notons également qu'il est possible pour un concepteur de créer son propre abstract à partir de son propre layout. Ceci peut être utilisé par exemple pour un circuit mixte, c'est à dire qui comporte une ou plusieurs parties analogiques conçues selon le flot "full custom" et une ou plusieurs parties numériques. La partie analogique peut donc tout à fait être intégrée au flot de placement et routage. C'est également la méthode utilisée lorsque le concepteur souhaite intégrer une mémoire RAM dans un circuit. Dans la plupart des cas le fondeur ne fournit pas le layout complet de la

mémoire mais seulement une vue "abstract".

L'outil de placement et de routage utilise des fichiers technologiques .lef, pour Library Exchange Format. Il comporte la liste de toutes les cellules avec la description de toutes les zones comportant du métal dans chaque cellule. Cela permet à l'outil de routage de ne pas faire de court-circuit lors du routage des signaux. Ils sont créés à partir de la vue "abstract" elle-même créée à partir du layout. En principe, les fondeurs fournissent un fichier .lef par bibliothèque. La figure 2.21 montre le résultat d'un circuit placé et routé automatiquement.

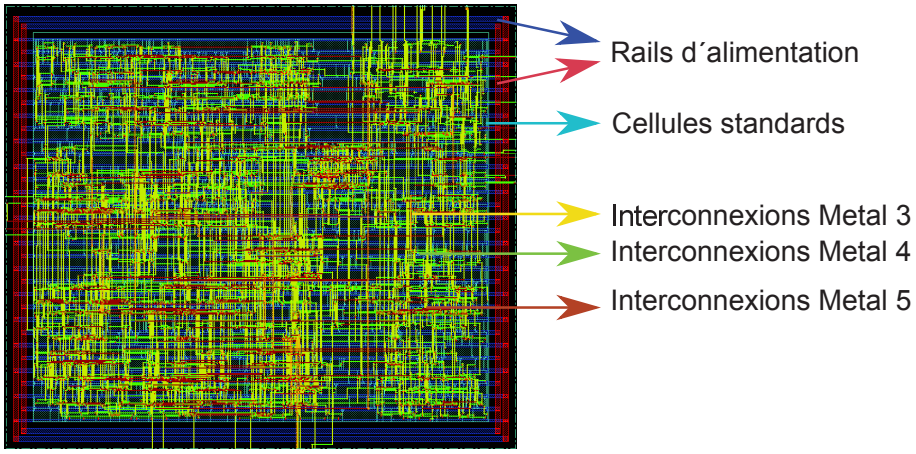


FIG. 2.21 – *Placement et routage complet d'un circuit numérique*

Une fois cette étape réalisée et terminée, la base de données du dessin des masques doit être extraite en vue de la fabrication. Cela consiste à générer un fichier GDSII, Graphic Database System II, depuis l'outil de placement routage, pour pouvoir l'importer dans un éditeur de layout, afin de faire les vérifications finales à savoir le DRC est le LVS. Ces 2 étapes ne seront pas décrites dans cette partie car elles l'ont déjà été dans la première partie de ce chapitre flot "full Custom", et qu'elles ne sont pas très différentes pour le flot numérique.

2.3.6 Simulation post-layout

Pour des circuits très complexes et avec des contraintes très fortes sur la fréquence d'horloge, les simulations numériques basées sur la caractérisation des cellules ne sont pas toujours suffisantes. Afin de prendre en compte les effets parasites du circuit une fois fabriqué, il est courant de faire des simulations complémentaires dites post-layout. Ces parasites augmentent le délai de propagation de tous les signaux. Si les temps de propagation deviennent trop longs et que le chemin critique devient si important qu'il n'y a plus de slack, alors le circuit ne fonctionnera plus, ou alors à

fréquence plus basse et le cahier des charges ne serait plus respecté. C'est pourquoi, si les spécifications sont très contraignantes, la simulation post-layout peut être très importante. Pour extraire ces parasites, les outils de conception utilisent des fichiers technologiques préalablement définis par la fonderie.

2.4 Conclusion

A notre époque les systèmes électroniques sont omniprésents, derrière lesquels se cachent une multitude d'étapes, depuis l'idée de l'application jusqu'à sa fabrication et sa commercialisation. Ce chapitre permet de situer la phase de conception dans ce processus si long et si complexe, depuis la spécification jusqu'à son implémentation au niveau physique. Ce chapitre permet également de montrer les 2 grandes familles de circuits intégrés, analogique (ou full custom) et numérique, pour lesquels les flots de conception sont très différents, ainsi que les contraintes et les outils de conception CAO. La figure 2.22 illustre une vue d'ensemble des 2 flots de conception décrit dans ce chapitre.

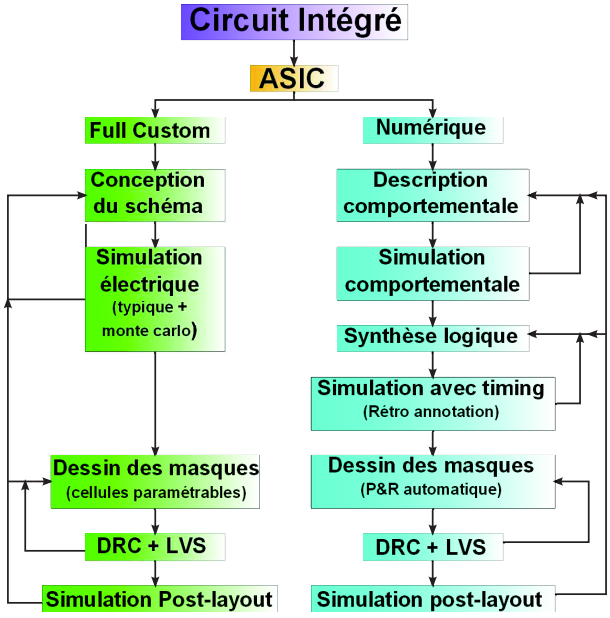


FIG. 2.22 – *Flot de conception d'un ASIC*

Enfin, ce chapitre permet d'introduire toutes les notions nécessaires qui permettront d'appréhender le chapitre "Kit de conception pour technologie hybride CMOS/Magnétique" que nous proposons, qui permet la conception de circuits intégrés "full custom" et numériques sur une telle technologie intégrant des composants magnétiques tels que des jonctions tunnel.

Chapitre 3

Conception de cellules innovantes non volatiles

3.1 Introduction

L'intérêt pour les technologies à mémoire émergentes est de plus en plus fort, et les efforts faits dans le domaine de la recherche de plus en plus grands. A travers deux projets ANR, nous avons eu l'opportunité de travailler sur le procédé TAS-MRAM de Crocus Technology [76]. Dans le premier projet, CILOMAG [8] pour CIrcuit LOgiques MAgnétiques, nos travaux portaient sur les aspects kit de conception, largement décrits dans le chapitre suivant. Dans le second projet, SPIN [9] pour SPintronics for Innovative Nanotechnologies, nos travaux portaient sur la conception de cellules innovantes, ce qui fait l'objet de ce chapitre.

Dans le but d'intégrer des composants magnétiques du type jonctions tunnel magnétiques à la circuiterie CMOS pour améliorer les performances et offrir d'éventuelles nouvelles fonctionnalités, notre approche a été de premièrement développer de nouvelles structures puis de les intégrer dans un flot de conception standard, sur des outils standards. Nous présentons dans ce chapitre une structure innovante d'un point mémoire non volatil que nous proposons, ainsi que son intégration dans une bascule de type flip-flop, en vue d'applications numériques non volatiles. Ces travaux ont été réalisés en utilisant le kit de conception que nous proposons dans le chapitre suivant.

Dans ce présent chapitre, nous commençons par présenter et rappeler l'architecture d'une mémoire SRAM volatile à 6 transistors, puis celle d'une SRAM non volatile, appelée "Black and Das" en lien avec ses inventeurs. Ensuite nous présentons l'architecture d'une cellule SRAM volatile à 4 transistors, puis celle de son homologue

sans résistance de charge, pour enfin présenter l'architecture que nous proposons. Il s'agit d'une cellule SRAM ultra compacte à 4 transistors non volatile et sans résistance de charge, intégrant des jonctions tunnel magnétiques, pour lesquelles nous présentons les différentes phases de maintien, de restauration et d'écriture selon différents mécanismes.

3.2 Cellule SRAM volatile à 6 transistors

La cellule SRAM à 6 transistors est de nos jours la cellule référence utilisée dans tous les systèmes intégrant de la mémoire SRAM. Comme présentée dans le chapitre 1, elle a beaucoup d'avantages tels que la rapidité, une faible consommation statique et une grande stabilité des niveaux logiques. En revanche, elle est peu dense mais elle est surtout volatile, donc ne peut pas répondre à un certain nombre de besoins. C'est pourquoi beaucoup d'efforts sont faits dans ce domaine pour tenter de remplacer à plusieurs niveaux les mémoires SRAM par des mémoires MRAM. Cependant, rendre une cellule SRAM 6T non volatile n'est pas impossible, et c'est ce que nous présentons dans le paragraphe ci-dessous. Nous rappelons tout de même dans ce chapitre l'architecture de cette cellule au niveau transistors (figure 3.1), composée de 6 transistors: 2 transistors d'accès et 4 transistors de mémorisation.

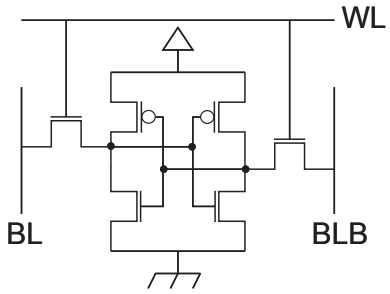


FIG. 3.1 – *Cellule mémoire SRAM volatile à 6 transistors*

3.3 Cellule SRAM non volatile "Black and Das"

Plutôt que de remplacer toute une mémoire volatile par une autre mémoire non volatile, flash par exemple, pourquoi ne pas transformer le point mémoire lui-même pour lui apporter de la non volatilité? C'est l'approche qu'ont eu W.C.Black et B. Das [21]. Leur structure, présentée en figure 3.2 dans sa version TAS, intègre d'une part les 6 transistors de la cellules SRAM 6T classique en noir sur notre illustration,

d'autre part 2 jonctions tunnel magnétiques pour apporter de la non volatilité, en rouge, ainsi que 2 transistors d'écriture des jonctions, en bleu.

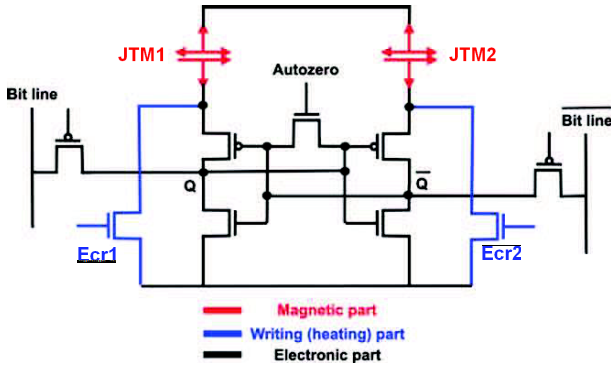


FIG. 3.2 – *Cellule SRAM 6T non volatile de type Black and Das*

Le mode de fonctionnement de cette structure, basé sur la technologie TAS-MRAM de la société Crocus Technology, est le suivant: les 6 transistors formant la SRAM fonctionnent exactement de la même façon que son homologue volatile, avec les mêmes performances en termes de vitesse et de consommation dynamique. Lorsque l'on souhaite sauvegarder l'information logique électrique dans la partie magnétique, alors il suffit d'activer les 2 transistors d'écriture Ecr1 et Ecr2 pour autoriser un courant à circuler à travers les jonctions tunnel. La piste de métal permettant de générer le champ magnétique servant à l'écriture des jonctions est commune aux 2 JTM, non représentée sur cette figure. L'étape d'écriture se fait donc en 2 phases: un premier MOS d'écriture, Ecr1 par exemple, est fermé seulement, ce qui permet de chauffer une première jonction grâce au courant la traversant, puis une aimantation disons dans le sens parallèle est imposée sur la couche de stockage de la jonction tunnel JTM1, grâce au courant circulant dans la ligne de champ d'écriture. Après un certain temps nécessaire au refroidissement pour assurer l'état codé, vient la seconde phase: seul le second MOS d'écriture Ecr2 est activé, ce qui permet de chauffer la deuxième jonction JTM2 par le courant la traversant. Une aimantation dans le sens antiparallèle est cette fois-ci imposée sur sa couche de stockage de la deuxième jonction, grâce au courant circulant dans la ligne de champ d'écriture, mais dans le sens inverse par rapport à la phase précédente. En effet, dans ce type de structure différentielle, composée de 2 branches symétriques, les 2 jonctions doivent systématiquement être dans deux états opposés. L'une a sa couche de stockage aimantée dans le sens parallèle donc présente une résistance faible, et l'autre a sa couche de stockage aimantée dans le sens antiparallèle et présente une résistance forte. Ceci est indispensable pour la phase de lecture qui est différentielle.

Cette cellule a à priori un objectif orienté plutôt mémoire, car il s'agit d'un point mémoire comme on les retrouve sous forme de matrice dans une SRAM classique. Il a été proposé en 2007 une bascule non volatile [129] [130] utilisant cette cellule "Black and Das", avec un objectif certainement plus orienté circuit intégré du type ASIC. La mémorisation locale peut être encore plus proche du calcul. Il a été proposé par l'université de Tohoku une cellule combinatoire "Full Adder" intégrant 4 jonctions tunnel magnétiques, 2 d'entre elles étant utilisées pour la partie "addition", et les 2 autres pour la partie "retenue" [86][84]. Le point fort illustré dans ces publications est la très faible consommation statique. On parle alors de concept de "Logic In Memory".

3.4 Cellule SRAM volatile à 4 transistors

Les premières cellules SRAM utilisées dans les systèmes électroniques étaient des SRAM à 6 transistors, puis est apparue la SRAM à 4 transistors, principalement appréciée pour sa taille réduite d'environ 30% par rapport à la SRAM 6T [96]. En effet, comme illustré sur la figure 3.3 (a) cette structure comporte 2 transistors de moins, et 2 résistances de plus, R1 et R2.

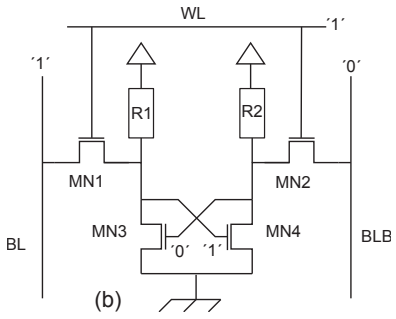


FIG. 3.3 – *SRAM volatile à 4 transistors classique*

Ces résistances dites de charges, fabriquées en même temps que les composants CMOS, en polysilicium, sont placées au-dessus des 4 transistors. Un gain en surface est obtenu lorsque la surface occupée par les transistors est supérieure à celle des résistances. Comme illustré sur la figure 3.3 (b) l'écriture de cette cellule se fait en préchargeant les bit lines BL et BLB respectivement à Vdd et à Gnd (ou inversement selon l'information à coder). Lorsque la commande des grilles des transistors MN1 et MN2 est activée par le signal word line WL, les 2 transistors d'accès MN3 et MN4 sont soit dans un état passant, soit dans un état bloqué en fonction de BL et BLB. Le latch va donc s'équilibrer en ayant 1 transistor MN4 passant avec Vdd sur sa

grille et l'autre MN3 bloqué avec Gnd sur sa grille. Lorsque la commande WL des 2 transistors MN1 et MN2 est désactivée, le latch sera alors dans un état stable. Cette stabilité est assurée grâce aux 2 résistances de charges. En effet, dans cette structure un des 2 transistors MN3 ou MN4 est systématiquement passant, ce qui permet à un courant de circuler en permanence dans une des 2 branches via R1 ou R2 et donc de maintenir un niveau logique assez élevé sur l'une des deux grilles de MN3 ou MN4 pour qu'un de ces transistors soit passant.

Bien qu'en termes de surface la SRAM 4T présente un avantage, elle est tout de même 4 fois plus encombrante qu'une DRAM à génération équivalente. De plus le courant permanent circulant dans une des 2 branches est un inconvénient fort du point de vue de la consommation statique, parmi d'autres. Pour remédier à ce phénomène de surconsommation, l'idéal serait d'avoir des résistances suffisamment élevées, de l'ordre du GigaOhm pour réduire drastiquement ce courant. Ceci reste toutefois une difficulté majeure du point de vue de la fabrication et la miniaturisation devient impossible à partir des procédés de fabrication inférieurs à 0.8 μm [96]. Le fait d'avoir une résistance si élevée rend cette structure SRAM 4T sensible au bruit. De plus, elle n'est pas aussi rapide que la SRAM 6T, donc son intérêt est largement moindre. C'est pourquoi cette architecture n'est plus utilisée dans les systèmes récents.

3.5 Cellule SRAM volatile loadless à 4 transistors

Comme nous venons de le présenter, la structure de la SRAM 4T avec résistance de charge a été abandonnée dans les années 90 du fait de l'impossibilité de fabriquer, sur une surface raisonnable, des résistances suffisamment élevées pour réduire le courant statique. En 2000, une nouvelle structure de SRAM 4T sans résistance de charge a été proposée [113] [94] [125]. On peut effectivement voir sur la figure 3.4 qu'il n'est plus nécessaire d'avoir de résistance de charge et que cette structure est composée uniquement de 4 transistors MOS, parmi lesquels 2 sont des NMOS et 2 sont des PMOS, alors que la précédente avec charge n'utilise que des NMOS.

Le mode de fonctionnement de cette cellule est le même que celui de la SRAM 4T classique. L'innovation de cette structure se trouve au niveau de la diversité des transistors MOS. Les 2 NMOS MN1 et MN2 sont identiques à la version avec charge et leur comportement l'est aussi, en revanche les 2 résistances de charge ont été remplacées par des transistors PMOS MP1 et MP2. Ces transistors servent de transistors d'accès lors des phases de lecture ou d'écriture, et de transistors de charge pendant la phase de maintien. Ces derniers doivent avoir une tension de seuil (Vt)

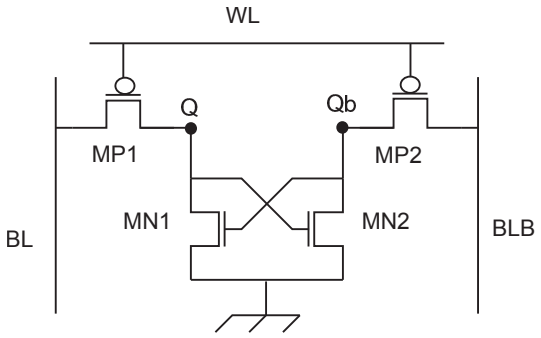


FIG. 3.4 – *SRAM 4T loadless volatile: 4 transistors sans résistance de charge*

plus faible que celle des NMOS, c'est à dire avoir un courant de fuite Ioff supérieur. Le seuil de tension des transistors étant inversement proportionnel au courant de fuite source - drain. Comme on peut le comprendre sur l'*equation 3.1* [100], plus le Vt est élevé, plus le courant de fuite sera faible, mais plus le transistor sera lent. A l'inverse, plus le Vt est faible, plus le courant de fuite est élevé, mais plus le MOS sera rapide.

$$Ids_{sub} \;=\; Kx \times \frac{W}{L} \times e^{\frac{Vgs}{nVt}} \times (1 \;-\; e^{\frac{-Vds}{Vt}}) \tag{3.1}$$

avec:

Kx: paramètre dépendant du procédé (typiquement 50 uA / V^2)
W: longueur du canal du transistor
L: largeur du canal du transistor
Vgs: tension grille / source du transistor
n = 1,5
Vt: tension de seuil du transistor

Généralement, dans les technologies qui offrent plusieurs gammes de transistors, les "High Vt" sont utilisés pour les applications à faible consommation, et les "Low Vt" sont utilisés pour des applications hautes fréquences. Ceci est de plus en plus vrai avec les technologies très submicroniques, qui sont utilisées pour des systèmes complet type SoC [50] qui intègrent sur le même circuit des blocs rapides fonctionnant à haute fréquence et des blocs basse consommation.

Le courant de fuite des PMOS étant donc supérieur à celui des NMOS dans cette structure, le niveau logique aux noeuds Q et Qb du point mémoire sera suffisamment élevé pour maintenir passant un des NMOS qui est déjà dans cet état à la suite d'une phase d'écriture. Par conséquent, l'autre NMOS est bloqué. Pendant la phase

de maintien, la structure SRAM 4T loadless est donc stable grâce à un courant de fuite. Les phases d'écriture et de lecture restent les mêmes, si ce n'est que la commande des grilles issues du signal Word Line doit avoir un niveau logique opposé à celui de la SRAM classique, car ce signal commande cette fois-ci des transistors PMOS.

Une étude a montré en 2009 qu'une mémoire matricielle à base de SRAM 4T loadless est plus avantageuse qu'une mémoire SRAM 6T. Que ce soit en termes de surface ou de consommation, le gain est notable, et sans perte de performance en vitesse [102].

3.6 Cellule SRAM non volatile loadless à 4 transistors

L'architecture que nous proposons ici a toujours le même objectif, c'est à dire utiliser la non volatilité des jonctions tunnel dans les circuits microélectroniques, avec pour but d'améliorer les performances électriques et d'offrir de nouvelles fonctionnalités. Le but est d'apporter des avantages par rapports aux architectures existantes sans faire de compromis sur les performances. Les critères visés sont tant au niveau de la taille de l'architecture, que de sa consommation, que de ses performances en vitesse. Pour cela, nous avons intégré des composants magnétiques dans des structures compactes, telle que la SRAM 4T. Ces composants sont des jonctions tunnel magnétiques pour lesquelles le mécanisme d'écriture est celui proposé par Crocus technology, à savoir TAS-MRAM. La raison principale de ce choix est que ces travaux ont été réalisés dans le cadre du projet de recherche ANR SPIN, dans lequel nous avions la possibilité de faire fabriquer un démonstrateur pour lequel la partie magnétique était assurée par Crocus Technology.

L'approche que nous avons eu pour apporter de la non volatilité aux circuits intégrés de façon générale a été sensiblement la même que W.C. Black et B. Das pour la SRAM 6T, mais en se basant sur la SRAM 4T loadless. La première architecture étudiée est celle présentée sur la figure 3.5 pour laquelle 2 jonctions ont été ajoutées en-dessous des transistors NMOS. Cette innovation a abouti à un brevet d'invention déposé en commun entre le LIRMM, CEA-Spintec et CMP [43]. Premièrement au niveau national en janvier 2011, puis au niveau international PCT en janvier 2012.

3.6.1 Phase de maintien et de restauration

Dans ce schéma, nous avons les 2 transistors PMOS MP1 et MP2 qui servent de charge pendant la phase de maintien, grâce à leur courant de fuite, à l'image du comportement de la cellule SRAM 4T loadless présentée précédemment. Le bit à

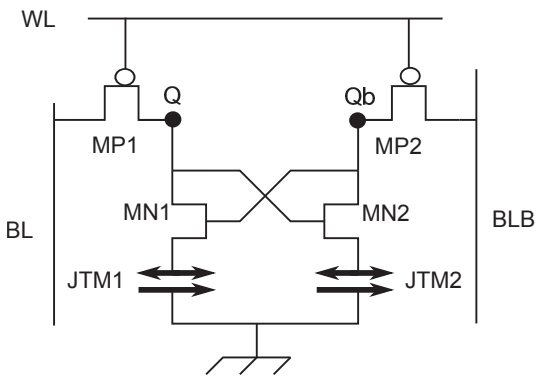


FIG. 3.5 – *SRAM 4T loadless non volatile V1*

mémoriser étant sauvegardé de façon différentielle, nous avons toujours une jonction à l'état parallèle et l'autre à l'état antiparallèle. La phase de restauration consiste à lire l'information contenue dans la partie magnétique et à l'écrire dans la partie électrique du latch. Pendant cette phase de restauration, ce montage agit comme un amplificateur de lecture, grâce aux 2 valeurs de résistances distinctes des JTM. Les 2 bit lines BL et BLB sont tout d'abord pré chargées à Vdd, puis la commande WL des transistors d'accès est activée. Ceci permet à 2 courants de circuler à travers les 2 NMOS MN1 et MN2 et donc à travers les 2 jonctions.

Dans l'hypothèse où les 2 résistances auraient la même valeur, le latch serait parfaitement symétrique. La tension aux noeuds Q et Qb serait identique, à savoir Vdd/2. C'est le cas de l'exemple des courbes noires sur la figure 3.6. Or dans le cas de structures comportant des JTM, les 2 résistances n'ont pas les mêmes valeurs car elles ne sont pas dans le même état magnétique. Le courant dans chaque branche est donc différent. On dit alors que le latch est déséquilibré, et qu'il se trouve dans un état métastable. C'est ce que nous illustrons à travers la figure 3.6 avec les courbes en rouge. A cet instant de la phase de lecture, la sortie Q a une valeur légèrement supérieure à Vdd/2 et la sortie complémentée Qb a une valeur légèrement inférieure à Vdd/2. C'est exactement pour cette raison là que le niveau de la TMR est important. Plus la TMR est élevée, plus le point de métastabilité sera loin de Vdd/2, aussi bien sur l'axe Q que sur l'axe Qb. La robustesse de ce type d'architecture dépend donc en partie de la TMR, de sa variation en fonction du procédé magnétique, ainsi que du dimensionnement des transistors.

Lorsque la commande WL est désactivée, le latch s'équilibre alors automatiquement. Les noeuds Q et Qb prennent alors la valeur la plus proche de celle de l'état métastable, Vdd pour l'un et Gnd pour l'autre. Dans l'exemple présenté ici, Q aura donc le niveau logique '1' et Qb le niveau logique '0'.

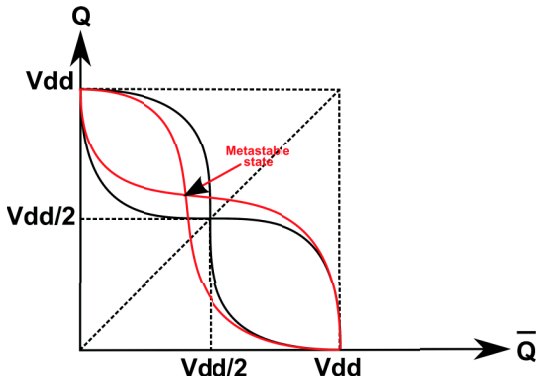


FIG. 3.6 – *Principe de déséquilibre du latch non volatile*

3.6.2 Phase d'écriture avec transistors de chauffe des JTM

Cette phase consiste soit à sauvegarder l'information électrique contenue dans le latch aux noeuds Q et Qb, soit à écrire une valeur quelconque imposée par l'ensemble du système, vers les jonctions tunnel. Cela peut être le cas pour une mémorisation de configuration par exemple. Nous proposons ici deux variantes de cette structure pour l'écriture.

La première variante consiste à ajouter 2 transistors d'écriture MP3 et MP4 au schéma de la figure 3.5, chacun connectés respectivement aux JTM1 et JTM2, comme illustré sur la figure 3.7.

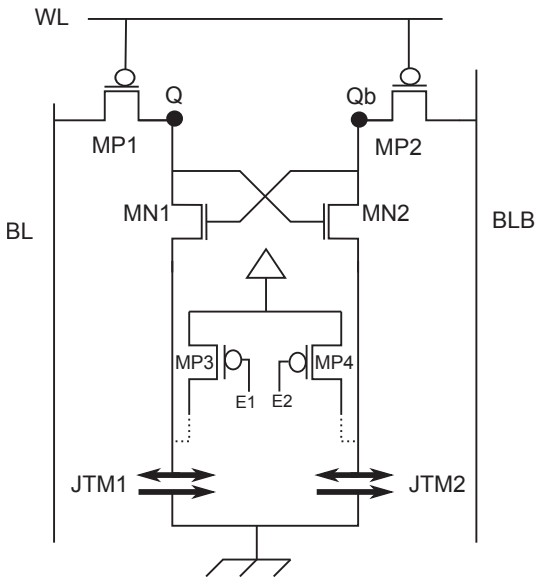


FIG. 3.7 – *SRAM 4T loadless non volatile avec transistors de chauffe des JTM*

Le chronogramme correspondant à ce fonctionnement est présenté sur la figure

3.8. La phase d'écriture se décompose ainsi: un premier PMOS, MP3 par exemple, est activé (t1) ce qui permet à un courant de circuler à travers la jonction JTM1 et par conséquent d'élever sa température au-delà de la température de blocage de la couche anti ferromagnétique. Ensuite, un courant générant le champ magnétique d'écriture traverse la piste métallique sous cette jonction JTM1 (t2), permettant ainsi d'imposer une aimantation sur la couche de référence entre t2 et t3. Suit la phase de refroidissement, toujours sous champ (t3), où le PMOS MP3 est à nouveau bloqué. La même séquence pour l'autre branche du latch est nécessaire, à savoir l'activation de l'autre PMOS MP4 (t5), permettant de chauffer la deuxième jonction JTM2, suivie de la génération du champ d'écriture dans le sens opposé (t6) pour ainsi imposer une aimantation contraire à celle de l'aimantation de la jonction JTM1 entre t6 et t7. Enfin, la seconde phase de refroidissement sous champ (t7) termine cette phase d'écriture.

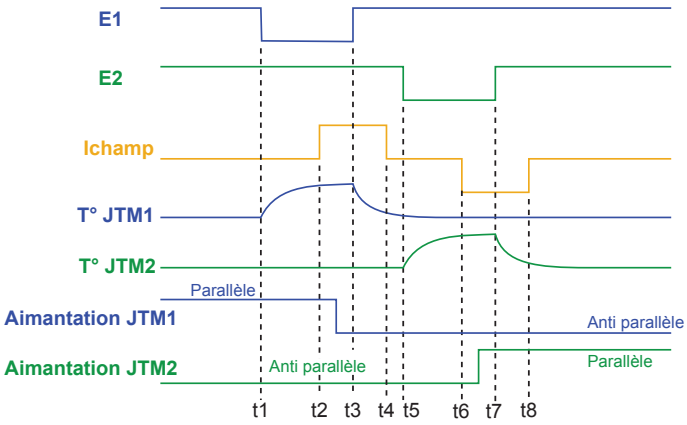


FIG. 3.8 – *Phase d'écriture de la cellule SRAM 4T loadless avec transistors de chauffe des JTM*

Cette variante de l'architecture est donc composée de 2 transistors PMOS d'accès MP1 et MP2, de 2 transistors NMOS de "pull down" MN1 et MN2, de 2 jonctions tunnel magnétiques et de 2 transistors d'écriture MP3 et MP4 permettant de chauffer les jonctions, le septième transistor Az étant optionnel.

Notons qu'il est tout a fait possible de remplacer tous les transistors NMOS en PMOS et tous les transistors PMOS en NMOS, celui de l'Auto zéro étant indépendant. Dans ce cas, les jonctions doivent être connectées à Vdd en série avec les PMOS, comme illustré sur la figure 3.9.

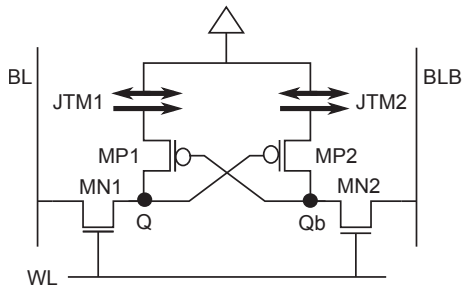


FIG. 3.9 – *SRAM 4T loadless non volatile V2*

3.6.3 Phase d'écriture sans transistors de chauffe des JTM

Dans l'architecture précédente où il est nécessaire d'intégrer des transistors de chauffe pour les jonctions tunnel, le gain en surface et en consommation statique est relativement intéressant, car le nombre total de transistors est de 6, contre 9 pour l'architecture " Black and Das". Néanmoins, nos réflexions se sont portées sur les transistors d'écriture MP3 et MP4 de la figure 3.7, avec l'objectif de réduire encore le nombre de transistors nécessaires. En effet, dans l'architecture que nous proposons sur la figure 3.10 ces 2 PMOS ont été supprimés.

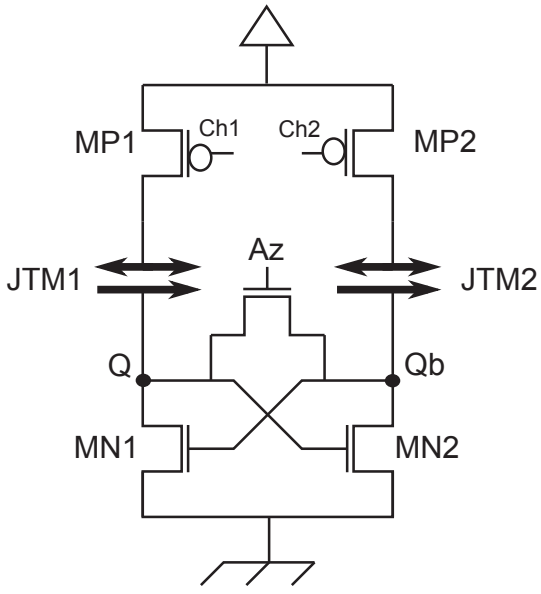


FIG. 3.10 – *SRAM 4T loadless non volatile compacte*

Cette cellule fonctionne de la même façon que son homologue pour la phase de maintien de la donnée, grâce à des courants de fuite à travers les PMOS, ainsi que pour la phase de restauration. L'innovation que nous proposons ici, qui a fait

l'objet d'un dépôt de brevet d'invention CEA-Spintec et CMP [97] en France en janvier 2011 puis à l'international en janvier 2012, porte sur la phase d'écriture des jonctions. Le mode de fonctionnement de cette phase, illustré sur le chronogramme de la figure 3.11 est le suivant: le principe est toujours de faire cette étape en 2 phases, afin d'écrire les 2 jonctions l'une après l'autre, car rappelons-le la ligne de champ d'écriture permettant d'imposer l'aimantation de la couche de stockage étant la même pour les 2 jonctions, il n'est pas possible d'écrire les 2 jonctions en même temps. Le courant doit donc circuler d'abord dans une branche du latch puis dans l'autre. Pour cela, le transistor MP1 est activé (t1), ce qui va permettre d'imposer un niveau logique '1' sur le noeud Q et donc d'activer le transistor NMOS MN2. Le transistor MP2 sera activé seulement après (t2), afin de permettre à un courant de circuler dans la branche de droite et ainsi chauffer la jonction JTM2, qui sera écrite par la ligne de champ (t3) lorsque la température aura dépassé la température de blocage entre t3 et t4. Ensuite, le signal Ch1 prend le niveau logique '1' pour bloquer le MOS MP1 (t4), alors que le signal Ch2 reste à '0'. Le MOS MP2 reste passant. Le noeud Qb prend alors la valeur '1' ce qui permet d'activer le NMOS MN1 cette fois ci. Dans certains cas de figure, essentiellement dépendant de la technologie CMOS, du type et de la taille des transistors, il se peut que la structure soit très stable. Dans ce cas, l'utilisation du transistor optionnel nommé Az, pour Auto zéro, est nécessaire. En effet, lorsque celui-ci est passant par application d'une tension positive sur sa grille, les 2 branches du latch sont court-circuitées, à la tension Vds près de ce MOS Az. Les noeuds Q et Qb sont alors au même potentiel ce qui permet de rendre le second transistor N passant. Le signal Ch1 peut alors reprendre la valeur '0' pour activer le PMOS MP1 (t6) et ainsi permettre à un courant de traverser la branche de gauche et chauffer la jonction JTM1, qui sera écrite par la ligne de champ (t7) lorsque la température de la jonction aura dépassé la température de blocage, entre t7 et t8. Les 2 commandes Ch1 et Ch2 peuvent alors être relaxées simultanément en t8.

Cette architecture innovante permet donc d'écrire les 2 jonctions en 2 cycles, l'une après l'autre, de la même façon que pour l'architecture de la première variante. Elle a l'avantage d'être plus compacte car elle n'est composée que de 4 transistors contre 6 pour la précédente, plus un optionnel, et sans compromis sur les performances en vitesse.

3.6.4 Dimensionnement et simulations électriques

Au cours de cette phase de conception de cellules innovantes compactes, plusieurs architectures ont été étudiées et simulées. Nous avons notamment essayé de placer

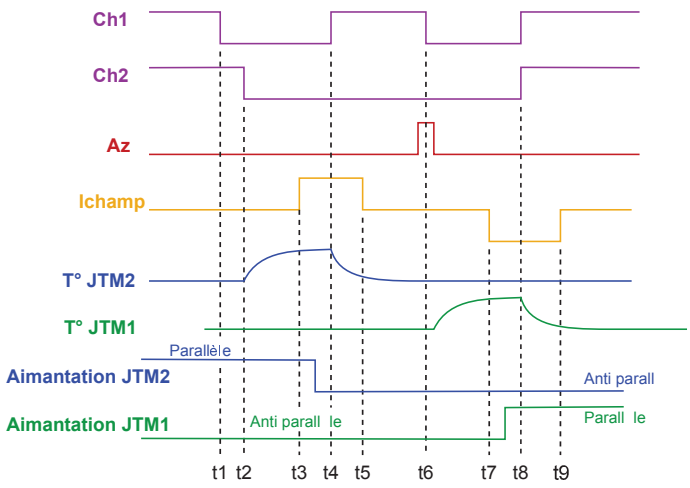


FIG. 3.11 – *Phase d'écriture de la cellule SRAM 4T loadless non volatile compacte*

les jonctions soit au-dessus des transistors PMOS soit au-dessous des transistors NMOS. Il s'est avéré que chacune de ces architectures fonctionnent, et de la même façon. Cependant, la configuration de la figure 3.10 est celle pour laquelle la taille des transistors est la plus petite. Cette configuration permet notamment d'avoir des transistors de taille minimum en technologie 65nm. Cela provient du fait que selon où sont placées les jonctions tunnel, les tensions Vgs et Vds aux bornes des transistors MOS ne sont pas les mêmes, ce qui influe sur leur comportement. Cependant, cette architecture nécessite de connecter les 2 bornes des JTM au niveau transistor. Comme les couches magnétiques sont réalisées au dessus de tous les niveaux d'interconnexion CMOS, cela nécessite deux empilements de VIAs pour chaque jonction, ce qui peut rendre le routage difficile dans le cas d'un système complexe. Dans le cas d'une cellule complexe intégrant des JTM, comme une bascule par exemple, ces empilements ne sont pas une contrainte très forte car la surface permet d'intégrer ces VIAs plus facilement.

Etudions le régime des transistors. Dans le cas d'un inverseur CMOS standard, les transistors sont, la plupart du temps, soit en régime bloqué soit en régime ohmique, appelé aussi régime linéaire. Prenons l'exemple du NMOS de l'inverseur de la figure 3.12:

Lorsque Vin = 0V, alors VgsN = 0V $<$ Vt. Le NMOS est donc bloqué quelle que soit la tension Vds, aucun courant Ids ne circule. Il s'agit du point A sur la figure 3.13. Lorsque Vin = Vdd, alors VgsN = Vdd $>$ Vt mais Vds = 0V. Il s'agit du point B de la figure 3.13. Quasiment aucun courant ne circule, le NMOS est en régime fortement ohmique. Ces deux régimes stables montrent que la consommation statique est très faible en technologie CMOS. En revanche, lors de la commutation

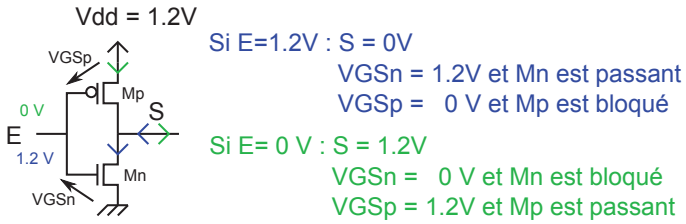


FIG. 3.12 – *Transistor MOS en petit signal*

du signal d'entrée, le NMOS passe dans un régime saturé pendant une courte durée. Il en est de même pour le PMOS. En effet, Vin passant de 0V à Vdd par exemple, on a VgsN = Vdd et VdsN = Vdd pendant ce laps de temps. Il s'agit du point C (régime saturé). Il y a donc un pic de courant Idssat entre le drain et la source, ce qui explique la forte consommation dynamique des circuits en technologies CMOS. Après un temps très court, Vds prend alors la valeur 0V et le NMOS se retrouve à nouveau dans un régime ohmique.

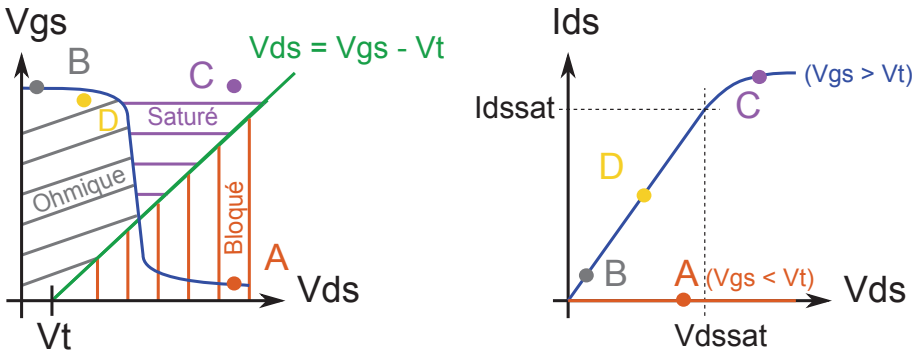


FIG. 3.13 – *Régime d'un transistor MOS*

Dans le cas de la structure que nous proposons, la phase d'écriture nécessite d'avoir un courant fort pour chauffer suffisamment les jonctions et dans un temps le plus court possible. L'idéal serait donc d'avoir les 2 transistors NMOS et PMOS d'une même branche tous les deux dans un état saturé en même temps, comme lors d'une phase de commutation, régime dans lequel le courant drain-source est le plus élevé. Cependant, les JTM étant résistives, le comportement des transistors est quelque peu différent d'une structure CMOS classique, même avec les Vgs = Vdd simultanément. En effet, lorsque qu'un courant circule dans une des branches du latch magnétique, il traverse une résistance série ce qui occasionne une tension non négligeable aux bornes de la jonction. Que la jonction soit dans un état parallèle ou antiparallèle, sa résistance est de l'ordre de quelques KOhms. La tension Vds des

transistors n'est donc plus maximale comme lors d'une phase de commutation. La figure 3.14 montre que lors de la phase d'écriture de la jonction JTM2, les transistors MP2 et MN2 ne sont pas dans un régime saturé, mais plutôt dans un régime linéaire. Bien que Vgs = Vdd, soit la tension maximum, les tensions Vds des 2 MOS sont trop faibles pour générer un courant maximum. Il s'agit du point D de la figure 3.13.

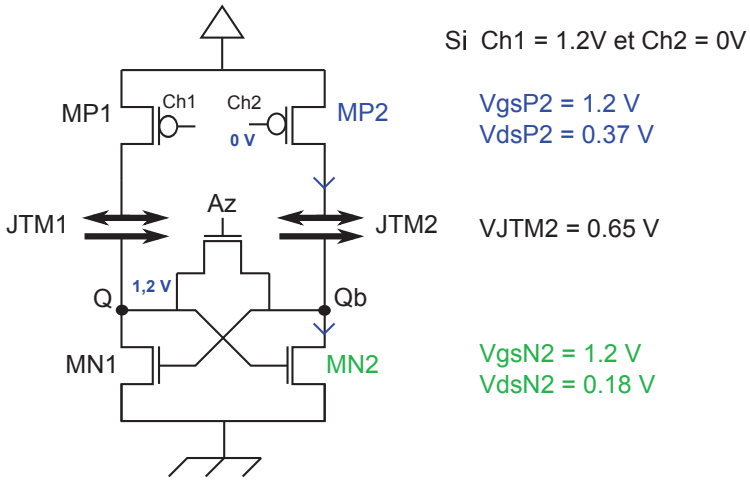


FIG. 3.14 – *Régime d'un transistor MOS pour une structure SRAM 4T loadless CMOS/Magnétique pendant une phase d'écriture*

Par ailleurs, lorsque les jonctions sont placées au-dessus des transistors P, les VdsP2 et VdsN2 restent les mêmes pour un courant de même valeur, mais la tension VgsP2 vaut alors 0.18V + 0.37V = 0.55V < Vdd. Ceci implique que la taille du transistor doit être plus importante pour générer le courant nécessaire au chauffage de la jonction. Le raisonnement est le même si les jonctions sont placées sous les transistors N. Les tensions Vds sont identiques pour un courant donné, par contre le VgsN2 vaut alors 1.2V - 0.65V = 0.55V. De même, il faudrait augmenter la taille de ce transistor pour générer le même courant de chauffe. La configuration retenue avec les jonctions entre les transistors est donc optimale du point de vue de la taille du latch, ainsi que du point de vue de la consommation statique.

L'ensemble de ces résultats et conclusion sur le dimensionnement des transistors ont été établis à partir de simulations électriques, en utilisant le modèle compact de simulation des jonctions tunnel magnétiques développé par le laboratoire CEA-Spintec. Ce modèle est largement décrit dans le chapitre dédié au kit de conception. Il n'intégre pas la possibilité de faire des simulations "Monte Carlo" sur les jonctions, cependant les simulations Monte Carlo ont tout de même été faites au niveau des transistors, sur plusieurs centaines d'itérations. Afin de simuler les variations du

procédé magnétique, nous avons considéré que les variations pouvaient être importantes d'un wafer à un autre, voire d'une zone d'un wafer à une autre, mais pas d'une jonction à une autre localement. Dans cette hypothèse qui nous parait tout à fait cohérente, nous avons fait des simulations Monte Carlo toujours sur les transistors mais en changeant la taille des jonctions, entre 110nm et 130nm de diamètre, soit 10% de variation. La valeur typique étant de 120nm. Enfin, nous avons fait des simulations paramétriques sur les jonctions selon cette gamme de variation.

3.7 Cellule Flip-Flop innovante non volatile

Comme nous l'avons déjà décrit, les derniers travaux sur les technologies non volatiles ont à ce jour été principalement axés sur les mémoires. Un des objectifs de cette thèse est de permettre la conception de circuits intégrant des jonctions tunnel magnétiques dans les circuits full custom et numériques. Pour cela il est indispensable d'avoir à disposition des registres de type flip-flop, non volatils. Nous avons donc couplé le latch non volatil compact présenté ci-dessus à un latch SRAM classique à 6 transistors pour créer une bascule non volatile. Comme nous l'avons décrit précédemment, la phase de sauvegarde de la partie électrique du latch dans la partie magnétique doit suivre un séquencement très particulier à cette architecture. Celui-ci consiste entre autre à piloter les signaux de commande Ch1 et Ch2. Cependant, il n'est pas envisageable de générer ces 2 signaux de commande pour chacune des bascules. Ils doivent donc tout deux être uniques pour l'ensemble du circuit. Par conséquent, en fonction de la valeur à sauvegarder, "0" ou "1", soit la branche de droite, soit la branche de gauche du latch non volatil doit être commandée en premier, pour que les jonctions soient dans une configuration parallèle / antiparallèle ou alors antiparallèle / parallèle. Il est donc nécessaire pour cela d'intégrer un aiguillage des signaux Chauf1 et Chauf2 vers Ch1 et Ch2 ou vers Ch2 et Ch1. C'est l'objectif du module "gestion_commande", composé de 4 transistors de type "pass gate" intégrés à cette bascule. Le principe de fonctionnement de ce module "gestion_commande" est le suivant:
Si Q = 0, alors Qb = 1, donc Ch1 = Chauf1 et Ch2 = Chauf2
Si Q = 1, alors Qb = 0, donc Ch1 = Chauf2 et Ch2 = Chauf1

Ceci permet de générer 2 signaux uniques Chauf1 et Chauf2 définis sur le chronogramme de la figure 3.11 pour l'ensemble du circuit, qui seront alors tous les 2 connectés à toutes les bascules, la gestion étant interne à chaque flip-flop en fonction du niveau de la sortie Q et de son complément Qb. La figure 3.15 montre

l'architecture complète de cette cellule.

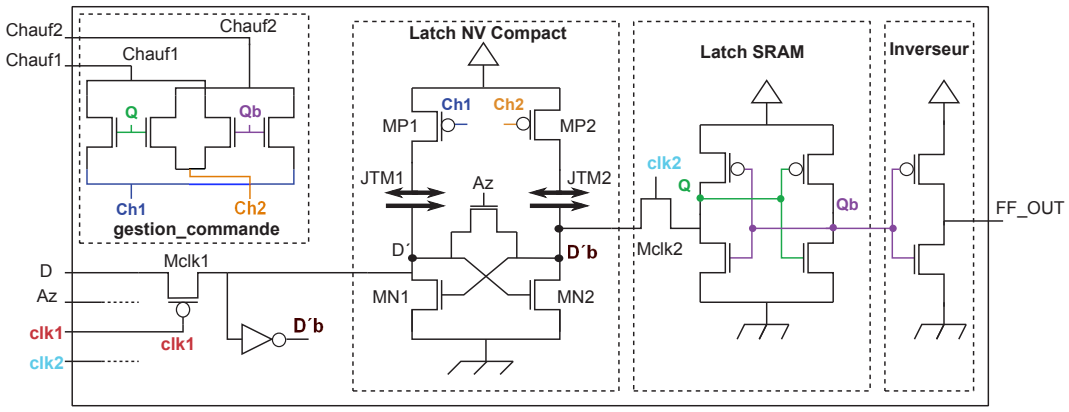


FIG. 3.15 – *Flip-Flop non volatile compacte*

On remarque également que cette flip-flop intègre un inverseur en sortie. Celui-ci a deux objectifs. D'une part il permet de regénérer le signal de sortie du latch volatil afin d'avoir un signal de sortie carré très bien défini et avec de bons signaux logiques. D'autre part, cet inverseur permet de proposer dans une bibliothèque plusieurs flip-flops basées sur le même fonctionnement, mais avec différentes sortances. En effet, plus l'inverseur de sortie aura de gros transistors, plus la flip-flop sera rapide et plus elle sera capable de charger toutes les capacités d'entrée des blocs avals connectés sur sa sortie FF_OUT.

Enfin, cette flip-flop n'a pas un seul signal d'horloge, mais 2 signaux, clk1 et clk2. Dans les bascules CMOS standards, seul 1 signal suffit car elles sont souvent composées de 2 latchs identiques, un actif sur niveau bas et l'autre sur niveau haut. Classiquement, lorsque le signal clk vaut '0' le premier latch, communément appelé master, prend la valeur de l'entrée D. Donc si D change alors la sortie de ce latch change aussi, mais pas celle de la bascule. En revanche, lors du front montant de clk, la valeur de D à cet instant est propagée dans le second latch, communément appelé slave. On a donc bien Q = D sur front montant de l'horloge. Dans le cas de la bascule non volatile que nous proposons, le principe est strictement le même en fonctionnement CMOS, c'est à dire l'un des deux transistors Mclk1 ou Mclk2 est passant et l'autre est bloqué, Mclk1 étant un PMOS et Mclk2 étant un NMOS. Par contre, lors d'une phase de restauration depuis la partie magnétique ou lors d'une phase de sauvegarde vers la partie magnétique, le latch non volatil doit être complètement isolé électriquement de la flip-flop, afin de ne pas perturber le fonctionnement global du circuit, car chacune de ces deux phases peuvent modifier temporairement la sortie du latch magnétique et donc celui de la flip-flop. Pour ce faire, les 2 transistors Mclk1 et

Mclk2 doivent être bloqués simultanément. Afin de ne pas avoir à gérer individuellement la commande de ces 2 transistors, nous avons conçu un module de génération de signaux qui permet de faciliter cette gestion. Celui-ci génère les signaux clk1 et clk2 à partir du signal d'horloge global du circuit et d'un signal "mag", de la façon décrite par les 2 tableaux de Karnaugh table 3.1 et table 3.2.

CLK1		clk
		0 | 1
Mag	0	0 | 1
	1	1 | 1

TAB. 3.1 – *Génération du signal clk1*

CLK2		clk
		0 | 1
Mag	0	0 | 1
	1	0 | 0

TAB. 3.2 – *Génération du signal clk2*

On déduit de ces 2 tableaux les 2 équations booléennes suivantes *equation 3.2* et *equation 3.3*:

$$\mathbf{clk1} \;=\; clk \;+\; mag \tag{3.2}$$

$$\mathbf{clk2} \;=\; clk \;.\; \overline{mag} \;=\; \overline{\overline{clk \;.\; \overline{mag}}} \;=\; \overline{\overline{clk} \;+\; mag} \tag{3.3}$$

Ces équations booléennes simplifiées permettent de construire ce module de génération de ces 2 signaux clk1 et clk2 pour un circuit full custom. Dans le cas d'un circuit numérique, il est tout à fait possible d'implémenter ce bloc en VHDL et de le synthétiser avec l'ensemble des autres fichiers source du design.

L'ensemble de ces modules ont été simulés séparément puis intégrés à la flip-flop. La bascule complète elle-même a été simulée électriquement, par simulations paramétriques sur les jonctions tunnel et Monte Carlo pour la partie CMOS, sur plusieurs centaines d'itérations. Ceci a permis de dimensionner cette cellule de façon à ce qu'elle soit la plus robuste possible vis à vis des variations, en vue de la fabrication d'un démonstrateur pour lequel le détail est abordé au cours du dernier chapitre.

3.8 Conclusion

Une grande majorité des recherches dans le domaine des technologies émergentes portent sur les aspects mémoires. Nous avons donc essayé de répondre à ce type de besoin en proposant une nouvelle architecture pour un latch SRAM non volatil compact, n'utilisant potentiellement que 4 transistors, voire 5 avec l'auto zéro optionnel. Cette structure est donc plus dense que les latchs des mémoires SRAM utilisés actuellement qui sont composés de 6 transistors, ce qui permet de répondre au besoin de densité d'intégration. Soit en réduisant la surface utilisée pour une capacité de stockage donnée, soit en augmentant la quantité d'informations pouvant être sauvegardées pour une surface donnée. De plus, cette structure est non volatile du fait qu'elle intègre des jonctions tunnel magnétiques, comme utilisées dans les mémoires MRAMs. Le mode de fonctionnement de la cellule SRAM à 4 transistors sans charge non volatile proposée étant le même que celui d'une SRAM à 6 transistors classiques, c'est à dire commander un point mémoire à partir d'une ligne Word Line et d'une colonne Bit Line, son intégration peut se faire sans difficulté. En revanche, il est possible de sauvegarder à n'importe quel moment les données électriques dans la partie magnétique, ou inversement de restaurer le contenu de la partie magnétique dans la partie électrique du latch. Enfin, cette nouvelle architecture permet d'avoir une consommation statique très faible car aucun courant ne circule pendant les phases de standby, hormis les courants de fuite.

Il est beaucoup moins courant de trouver des papiers de recherche au niveau des circuits intégrés non volatils, analogique ou numérique, que pour des mémoires. Nous pensons qu'il serait très intéressant d'apporter de la non volatilité dans les circuits numériques complexes et que ce serait un atout, tant du point de vue de la sécurité si les données sont sauvegardées très régulièrement, que du point de vue de la consommation en coupant l'alimentation des parties inactives. C'est pourquoi nous proposons également une cellule du type bascule D, communément appelée flip-flop, composée d'un latch non volatil et d'un latch SRAM classique. On peut donc tout à fait imaginer concevoir un circuit numérique dans lequel toutes les flip-flops ou une partie seulement seraient des bascules non volatiles. C'est ce que nous proposons dans le 5ème chapitre de ce manuscrit. Pour cela, il est indispensable d'avoir tous les outils de conception et un flot de conception numérique permettant d'intégrer des bascules non volatiles. C'est ce que nous proposons dans le chapitre suivant, qui décrit la mise en place de flots de conception et le développement d'un kit de conception complet full custom et numérique, permettant la conception de circuits intégrés complexes intégrant des jonctions tunnel magnétiques.

Chapitre 4

Kit de conception pour technologie hybride CMOS/Magnétique

4.1 Introduction

Que ce soit pour la conception de circuits analogiques ou numériques, les outils de conception ne sont pas les mêmes. Nous avons vu dans le chapitre décrivant ces flots de conception que les spécifications du cahier des charges sont différentes et que le mode de conception l'est aussi. Ceci implique que chaque flot de conception utilise des outils spécifiques chacun associés à des fichiers technologiques spécifiques. Un des objectifs de cette thèse est d'être capable de concevoir un ASIC aussi bien analogique que numérique sur une technologie hybride CMOS/Magnétique selon des flots standards, et surtout en utilisant les outils de conception industriels standard. Pour cela nous proposons un kit de conception complet pour lesquels des fichiers technologiques ont été développés pour l'ensemble des étapes de conception d'un ASIC analogique et numérique. De plus, du fait que la technologie intègre des jonctions tunnel magnétiques et qu'il faille générer et gérer certains signaux spécifiques pour l'écriture et la lecture, nous avons mis en place un flot de conception spécifique basé sur les flots de conception standard d'ASIC. Nous présentons donc tout d'abord dans ce chapitre le procédé hybride CMOS / Magnétique pour lequel le kit de conception a été développé ainsi que les différents modules développés et disponibles dans ce PDK (Process Design Kit). Ensuite nous abordons l'ensemble des étapes de la conception full custom de circuits, puis celles de la conception de circuits numériques. Enfin la dernière partie traite de la conception de générateurs de courant d'écriture permettant de générer le champ de retournement des jonctions en technologie TAS et l'intégration de ces générateurs dans le flot de conception numérique. Dans ce cha-

pitre, nous ne revenons pas sur les principes de base de la conception largement décrit dans un chapitre précédent. Par contre, nous détaillons l'ensemble des travaux qui ont été faits pour la mise en place de ce kit de conception ainsi que les différentes implémentations nécessaires relatives à l'utilisation de jonctions tunnel magnétiques.

4.2 Procédé hybride CMOS / Magnétique

Ce kit de conception a été développé dans le cadre de 2 projets de recherche ANR, CILOMAG [8] et SPIN [9]. Dans ces 2 projets, un des objectifs était de valider des architectures sur silicium. Il a donc été mis en place un flot de fabrication spécifique. La partie CMOS a été fabriquée chez le fondeur STMicroelectronics et la partie back end magnétique conjointement entre Crocus Technology et le CEA-LETI. Le procédé microélectronique est 130n HCMOS9GP, qui comporte 6 niveaux de métal, dont le dernier est très épais. Les contraintes pour le post-process magnétique TAS étaient telles que le dernier niveau de métal ne devait pas être épais, essentiellement pour la raison qu'il faut plus de courant pour générer le même champ. Il a donc été demandé à ST d'arrêter prématurément leur procédé de fabrication pour livrer spécifiquement au consortium des wafers pour lesquels le dernier niveau de métal était le "metal 5", et non pas le "metal 6". Ensuite, plusieurs niveaux de métal, d'interconnexions et de couches magnétiques ont été déposés. La figure 4.1 montre la vue de coupe de l'ensemble de ce procédé hybride.

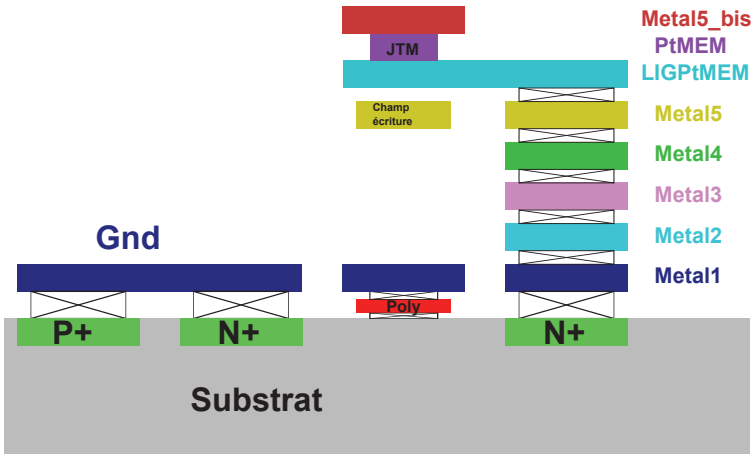


FIG. 4.1 – *Vue de coupe du procédé hybride CMOS / Magnétique TAS*

Afin de pouvoir concevoir des dispositifs intégrant des jonctions tunnel, il est indispensable de disposer de ces niveaux dans le kit de conception. L'ensemble de ces layers ont été implémentés dans les fichiers technologiques pour l'environnement

Cadence, notamment pour qu'ils apparaissent dans le LSW (Layer Selection Window) comme le montre la figure 4.2 illustrant le layout d'une jonction tunnel et de ses niveaux d'interconnexion.

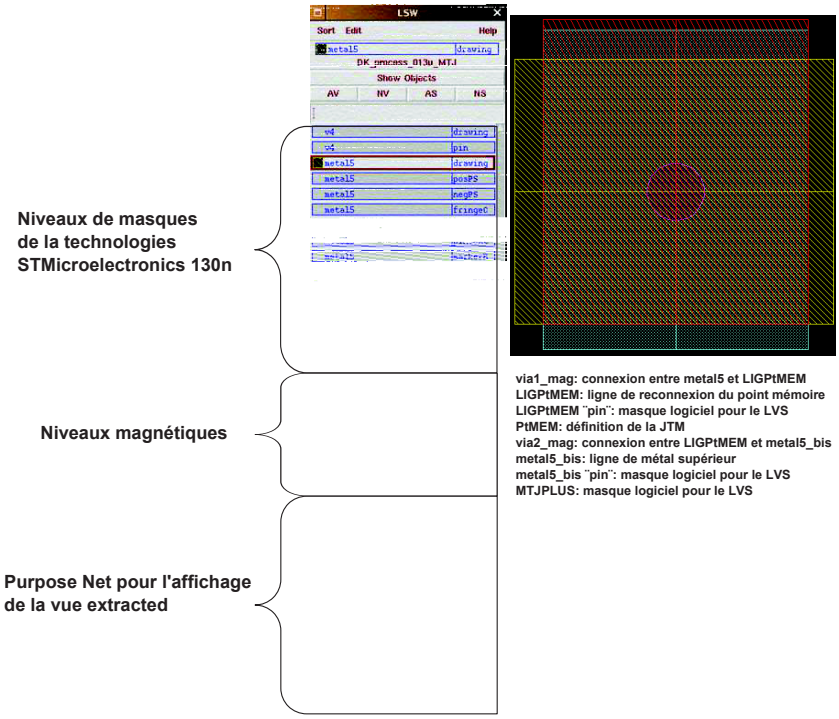


FIG. 4.2 – *Environnement graphique du procédé hybride sous Cadence*

Pour chacun de ces signaux nous avons défini d'une part l'apparence de l'affichage sur l'éditeur de dessin de masques, c'est à dire la couleur, l'aspect etc., ainsi qu'un numéro GDSII (Graphic Data System). Ces numéros sont utilisés pour exporter et importer des bases de données complètes de circuits intégrés, entre les différents partenaires des projets.

4.3 Schéma et modèle compact pour la simulation électrique de jonctions tunnel

Lors de la conception full custom d'un circuit, la toute première étape consiste à dessiner le schéma de l'architecture en important le symbole de chacun des composants à utiliser. Pour cela il est nécessaire d'avoir une vue "symbol" de la jonction tunnel. Celui que nous proposons comporte 6 terminaux:

– BL0 et BL1: ceux sont les 2 terminaux auxquels sont connectées les composants microélectroniques, la plupart du temps des transistors mais également des

capacités et ou des résistances, en fonction des besoins du montage.

- FL0 et FL1: ce sont les terminaux auxquels est connectée la ligne de champ d'écriture, dans laquelle un fort courant circule lors de la phase d'écriture.
- th: il s'agit d'un terminal utilisé seulement lors de la phase de conception "schematic" pour la simulation. Il permet d'extraire la température de la jonction.
- my: il s'agit également d'un terminal utilisé seulement lors de la phase de conception "schematic" pour la simulation. Il permet d'extraire l'état de l'aimantation de la jonction, parallèle ou antiparallèle.

La figure 4.3 montre le symbole d'une JTM dans un environnement Cadence / Schematic Composer, ainsi que le schéma du latch compact non volatile dans lequel 2 JTM sont utilisées.

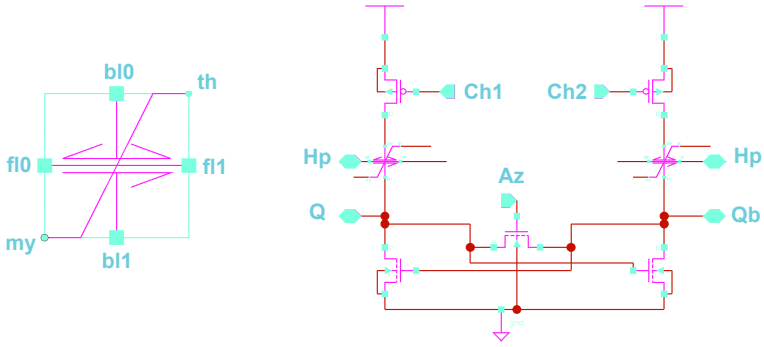


FIG. 4.3 – *Vue "symbol" d'une JTM et schéma intégrant des JTM*

L'étape suivante est la simulation. Au niveau full custom, les simulations sont dites électriques et sont basées sur des modèles compacts de simulation. Pour des transistors MOS, les modèles sont définis sous un standard "bsim". BSIM signifie Berkeley Short-channel IGFET Model, car en effet ce standard a été développé par l'université UC Berkeley [105]. Au fur et à mesure des avancées des technologies, la précision requise modélisant le comportement des transistors s'accroit. Depuis sa création, plusieurs versions standard ont été proposées: BSIM3 [58] très couramment utilisé dans les technologies les plus matures, BSIM4 [59] permettant de prendre en compte plus précisément les effets de courant de fuite des transistors ainsi que le comportement sous le seuil, BSIMSOI [61] pour les procédés SOI et dernièrement le standard BSIMCMG [60] le premier modèle standard pour les transistors a plusieurs grilles FinFET.

Nous proposons dans ce kit de conception un modèle compact pour les jonctions tunnel magnétiques. Celui-ci a été développé pour le simulateur "spectre" et permet de simuler électriquement une structure intégrant à la fois des transistors et des jonc-

tions tunnel. Les paramètres que l'on peut extraire de ces composants magnétiques à partir de ce modèle sont les suivants:

- La température de la jonction. En effet ce paramètre est primordial pour des simulations de jonctions TAS. Il est nécessaire de connaitre la température de la jonction pour d'une part appliquer le champ d'écriture au bon moment et d'autre part pour ne pas sur dimensionner les transistors au risque d'imposer aux jonctions une température supérieure à leur température de destruction. De plus, il est important de pouvoir contrôler également la phase de refroidissement afin d'appliquer le champ d'écriture pendant un temps suffisamment long pour s'assurer de la stabilité de la valeur écrite. Notons que le paramètre de température est également important en technologie STT, car la stabilité et le bruit thermique des jonctions y sont directement liés.
- L'aimantation de la couche de stockage. Ce paramètre permet de s'assurer de l'état de la jonction lors de l'écriture afin de vérifier si la donnée restaurée est conforme au fonctionnement souhaité ainsi que de vérifier le changement d'état lors de la phase d'écriture. Cela permet également de s'assurer que les 2 jonctions sont systématiquement dans 2 états magnétiques opposés l'une par rapport à l'autre.

La figure 4.4 montre un résultat de simulation du latch compact non volatil où l'on remarque le comportement dynamique des jonctions qui a été implémenté dans le modèle de simulation, aussi bien pour le paramètre de température que celui de l'aimantation. Notons que le champ d'écriture est maintenu pendant les 2 phases de refroidissement afin de stabiliser la valeur écrite.

Ce modèle de simulation a été développé par le laboratoire CEA-Spintec [88], [118]. C'est un atout considérable pour la phase de conception car cela permet de dimensionner les transistors précisément en fonction du cahier des charges et en fonction du procédé magnétique. Cela permet également d'évaluer la robustesse d'une architecture en faisant des simulations Monte Carlo sur les transistors, car le phénomène des variations de fabrication magnétique n'est pas encore implémenté dans le modèle à ce jour. Cependant, il est possible de changer la taille typique des jonctions et de faire des simulations Monte Carlo sur les transistors ou de faire des simulations paramétriques sur la taille des jonctions.

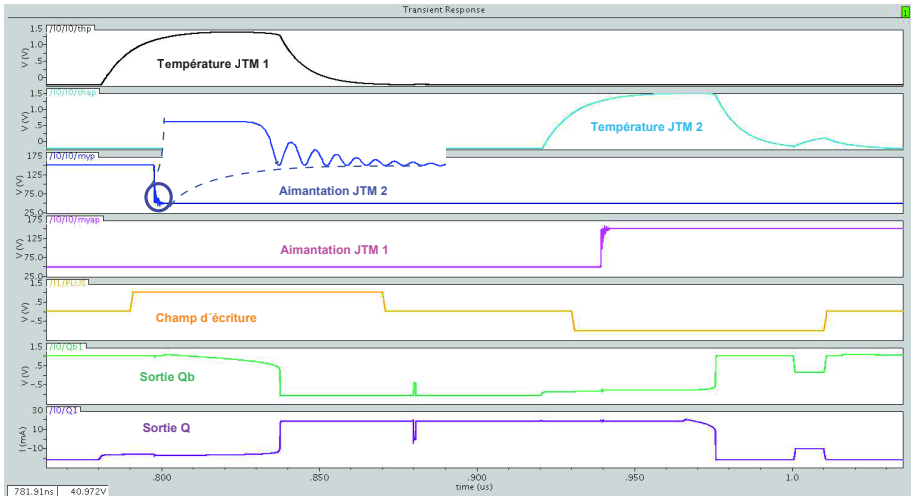


FIG. 4.4 – *Simulation électrique du latch compact non volatil à partir du modèle compact de jonctions TAS*

4.4 Cellule magnétique paramétrable: P-Cell

Le kit de conception que nous proposons intègre une cellule paramétrable d'une jonction tunnel. Cela permet d'ajuster les paramètres aussi bien pour la simulation que pour l'implémentation physique du dessin des masques. Les paramètres pouvant être modifiés par l'utilisateur sont les suivants:

- "a" et "b": ces paramètres définissent la taille de la jonction selon l'axe des X et l'axe des Y. Si ces 2 valeurs sont identiques alors la jonction est ronde, si elles sont différentes alors la jonction est ovale. Le layout de la JTM est alors automatiquement généré avec les tailles imposées dans les propriétés des jonctions, chacune pouvant avoir des valeurs différentes.
- MET5_bis_width: il s'agit de la largeur de la piste de métal permettant de connecter la jonction par son terminal supérieur.
- LIGPtMEM_width: il s'agit de la largeur de la piste de métal permettant de connecter la jonction par son terminal inférieur.
- MET5_width: il s'agit de la largeur de la piste de métal de champ d'écriture, placée au-dessous des jonctions.
- Un ensemble de paramètres technologiques directement liés au procédé de fabrication magnétique. Ceux-ci sont utilisés par le modèle de simulation uniquement et influent sur le comportement des jonctions seulement.

L'édition de ces paramètres se fait de façon graphique comme le montre la figure 4.5, aussi bien sous Cadence / Schematic Composer pour la partie schéma que sous

Cadence / Virtuoso pour la partie dessin des masques.

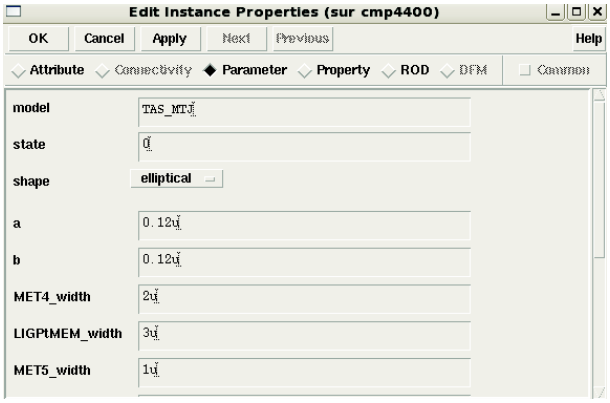


FIG. 4.5 – *Cellule paramétrable d'une jonction tunnel magnétique sous Cadence*

Cette cellule paramétrable permet de générer le dessin des masques automatiquement, en respectant l'ensemble des règles de dessin aussi bien du CMOS que du magnétique, de même que toutes les cellules paramétrables fournies par les fondeurs dans les kits de conception.

Cette P-Cell a été développée en langage Skill. Ce langage de développement est propre à l'outil Cadence et a été initialement créé à l'université UC Berkeley, soumis à une conférence IEEE en 1990 [13]. Son nom provient originalement de "Silicon Compiler Interface Language" (SCIL), prononcé "SKIL" [47]. Ce type de langage, qui s'apparente au langage C, permet d'utiliser des procédures intégrant des paramètres qui peuvent être par exemple des règles DRC qui définissent une certaine distance entre 2 niveaux de métal. Ce programme est ensuite chargé dans une vue layout et sera sauvegardé en cache pour être exécuté à chaque utilisation de la P-Cell dans une architecture.

4.5 Règles de dessin de la technologie CMOS Magnétique

Afin d'assurer une bonne fabrication des composants magnétiques, Crocus Technology et le LETI, tous deux en charge du post-process magnétique, ont défini les règles de dessins à respecter. Ces règles sont du type espacement minimum, taille minimum ou fixe, encombrement d'un niveau par rapport à un autre, interdiction de dessiner certains niveaux seul par exemple. Toutes ces règles sont des règles classiques de la conception de circuits microélectroniques. En revanche, une particularité aux technologies CMOS / Magnétique TAS est le sens du courant dans la ligne d'écriture. L'aimantation de la couche de stockage d'une jonction tunnel est modifiable

par l'utilisateur. En revanche, l'aimantation de la couche de référence est fixe. Elle est imposée aux jonctions par le fondeur lors de la fabrication en appliquant une aimantation dans un sens défini identique à la couche anti ferromagnétique placée sous la couche de référence. Cette étape se faisant sur l'ensemble du wafer, toutes les jonctions sont polarisées dans le même sens. Il est donc indispensable que toutes les jonctions de tous les circuits étant fabriqués sur un même wafer soient orientées dans le même sens, afin d'avoir une ligne de courant d'écriture perpendiculaire à l'aimantation de la couche de référence. C'est pourquoi nous avons implémenté dans le fichier technologique de vérification des règles de dessins, pour l'environnement Cadence / Assura, une section qui permet d'identifier l'orientation des jonctions sur le dessin des masques. Si une jonction n'a pas la bonne orientation, alors une erreur DRC apparaît afin d'alerter le concepteur pour qu'il modifie son layout en conséquence. Pour ce faire, un niveau MTJPLUS a été ajouté à la liste des layers. Le DRC vérifie que ce polygone soit positionné de façon verticale comme illustré sur la figure 4.6. Ce layer n'est pas utilisé pour la fabrication des circuits et n'engendre pas la fabrication d'un masque spécifique. Il est utilisé seulement par les outils de vérification. On dit que c'est un layer CAD (Computer-Aided Design).

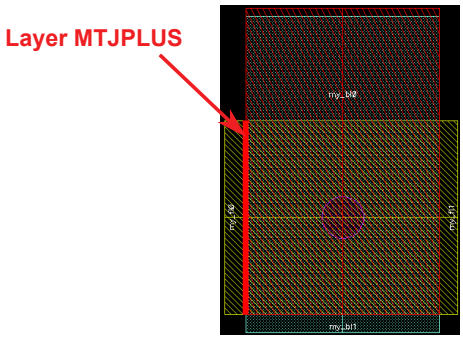


FIG. 4.6 – *Détection de l'orientation des jonctions tunnel magnétiques par le DRC*

De plus, afin de pouvoir vérifier les règles de dessin de la partie CMOS indépendamment de la partie magnétique, un switch a été ajouté dans le but de pouvoir faire un DRC spécifique ou global sur la totalité du layout. Cela facilite la conception et l'analyse des rapports DRC. Enfin, la taille de la jonction étant imposée par les partenaires en charge de la fabrication de la partie magnétique, cette taille fixe de diamètre de 120n est vérifiée lors du DRC.

4.6 Extraction des JTM - LVS mixte CMOS Magnétique

Bien qu'il y ait des outils de génération automatique de layout à partir du schéma, le risque d'erreur lors du placement des composants et lors du routage de ces composants est très important. C'est pourquoi l'étape de vérification de conformité entre le schéma et le layout, le LVS pour "Layout Versus Schematic", est une étape cruciale dans la conception full custom. Afin de pouvoir faire un LVS sur un circuit composé à la fois de composants de base de la microélectronique et des jonctions tunnel magnétiques, tous les composants doivent être reconnus. L'étape de vérification LVS nécessite d'une part d'avoir la "netlist" issue du schéma et d'autre part celle du dessin des masques.

En ce qui concerne celle issue du schéma, elle est générée à partir de la vue "schematic" dans lequel sont instanciés tous les composants. Néanmoins, chacun d'entre eux doit avoir une vue "auLVS", qui est utilisée par l'outil de génération de "netlist". Pour les jonctions tunnel, d'une part cette vue doit être créée, mais l'ensemble de ses éléments devant être extraits dans le fichier de liste des noeuds doivent être définis dans le cdf (Component Description Form) du composant. A savoir, l'ordre des terminaux, le nom du composant et de son modèle entre autres.

En ce qui concerne la "netlist" issue du layout, elle doit être extraite depuis le dessin des masques. L'outil d'extraction du layout se base sur des fichiers technologiques qui lui permettent de reconnaitre chacun des composants et de ses terminaux à partir des masques utilisés. Pour que l'outil reconnaisse les jonctions, plusieurs sections ont dues être implémentées dans ce fichier technologique:

- définition de tous les niveaux de masques relatifs au post-process magnétique, que ce soit les niveaux physiques d'implémentation ou les niveaux logiciels, comme le layer MTJPLUS par exemple qui permet d'identifier et de différencier le terminal Fl0 du terminal Fl1, ainsi que l'orientation de la jonction.
- reconnaissance de la jonction elle-même et définition de tous ses terminaux.
- reconnaissance de tous les terminaux d'une jonction tunnel: Bl0 et Bl1 pour la connexion aux composants, Fl0 et Fl1 pour les connexions aux générateurs de courant d'écriture.

L'exemple ci-dessous montre la façon dont a été implémentée la définition d'une jonction et celle de l'extraction de sa taille en langage Cadence/Assura, à partir de layers dérivés, eux-mêmes définis par la combinaison de layers physiques de fabrication.

smtj = geomOverlap(geomAndNot(geomEnclose(geomAnd(metal5 LIGPtMEM) PtMEM) nitride) MTJPLUS)

smtj_size = geomGetBBox(geomInside(PtMEM smtj))

Afin de pouvoir comparer les deux fichiers de description du circuit des vues "schematic" et "layout", l'outil de LVS s'appuie également sur un fichier de comparaison, dans lequel sont définis les composants qui doivent être vérifiés ainsi que les règles de mise en série et parallèle des composants. A partir de ces fichiers, il est donc possible de faire des vérifications LVS de circuits full custom hybrides CMOS / Magnétiques dans lequel des composants microélectroniques et magnétiques sont utilisés. Les 2 terminaux Fl0 et Fl1 sont également pris en compte dans cette vérification LVS, ce qui implique que cette étape permet aussi de s'assurer que le courant circulera dans les lignes de champ exactement dans le sens prédéfini par simulation du schéma. Ces travaux ont été développés pour l'environnement Cadence, pour les outils Schematic Composer, Virtuoso et Assura.

4.7 Extraction de composants parasites pour la simulation post-layout

La simulation post-layout peut s'avérer indispensable lorsque le cahier des charges de la conception est très contraint. Nous avons donc mis en place l'extraction de capacités parasites pour 2 outils de Cadence. D'une part Diva, le plus ancien des 2 qui est plutôt utilisé au niveau bloc, et Assura le plus récent des 2, très performant pour des architectures complexes. La raison de ce développement était de simplifier le flot de conception en permettant de faire toutes les étapes de conception à partir du même logiciel: Cadence. En effet, le flot préconisé et supporté par le fondeur STMicroelectronics est Cadence pour la conception full custom, Synopsys pour la synthèse, et Mentor Graphics / Calibre pour les vérifications DRC et LVS. Ceci impose la contrainte d'avoir à disposition toute cette suite logicielle et de maîtriser tous les outils.

La méthode d'extraction est différente entre Diva et Assura, et pas seulement du point de vue du langage. Pour Diva il est nécessaire de calculer chacun des composants à partir des paramètres technologiques disponibles dans le Design Rule Manual fourni par le fondeur. Ces paramètres sont l'épaisseur de chaque métal et donc de chaque diélectrique les séparant, ainsi que la résistivité des oxydes. La valeur d'une capacité planaire est définie par la relation de l'*equation 4.1.*

$$C = \epsilon_0.\epsilon_r.\frac{S}{e} \tag{4.1}$$

avec:

$\epsilon_0 = \frac{1}{36\pi.10^9}$: permittivité relative du vide (F/m)

$\epsilon_r = 8{,}84.10^{12}$: permittivité relative de l'isolant (sans unité)

S = Surface de la capacité (μm^2)

e = épaisseur entre les 2 armatures de la capacité ((μm))

Par ailleurs, bien que la capacité de surface soit l'élément le plus significatif, s'ajoute également à une capacité planaire la capacité de bord, appelée "fringe capacitance". Il est possible d'implémenter ces 2 éléments dans le langage Diva, ce qui permet d'apporter un maximum de précision aux valeurs de capacités parasites extraites par l'outil. En plus des capacités inter métal verticale, les capacités horizontales entre 2 pistes de même niveau, dites de "cross talk", doivent être également considérées pour que l'extraction soit la plus réaliste par rapport aux circuits fabriqués sur silicium. La relation de l'*equation 4.1* reste valable, si ce n'est que la surface devient la longueur de la piste multipliée par l'épaisseur du métal et l'épaisseur devient l'espacement entre les pistes. On peut donc définir chaque composant parasite par l'*equation 4.2* suivante:

$$C_{totale} = C_{surface} + C_{fringe} + C_{cross\_talk} \tag{4.2}$$

L'implémentation de l'extraction de capacités parasites consiste alors à définir dans le fichier technologique le coefficient de chaque composantes d'une capacité parasite (surface, fringe, cross talk) pour chacune d'entre elles du procédé de fabrication. L'ensemble des 35 capacités parasites du procédé STMicroelectronics 130n présentées sur la figure 2.12 ont donc été implémentées dans le fichier technologique Diva correspondant.

Notons que dans la liste des noeuds, ces composants peuvent apparaitre sous 2 conventions de référence, selon 2 switchs:

- extPAR_CapPairNode: toutes les capacités sont référencées par rapport aux noeuds qui la composent. Exemple: $C_{Vin-ref}$
- extPAR_CapSingleNode: toutes les capacités sont référencées par rapport au substrat. Exemple: $C_{Vin-Gnd}$

L'outil de référence préconisé par le fondeur ST pour l'extraction de composants parasites est l'outil StarRCXT de Synopsys. Nous avons donc validé notre implémentation en comparant les valeurs extraites par Diva et celle extraite par StarRCXT, et cela pour chacune des capacités parasites. Ceci a été fait à l'aide de plusieurs motifs de test pour chacune de ces capacités, de longueurs et de largeurs très différentes pour couvrir un maximum de cas possibles. Environ 200 motifs de test ont été né-

cessaires pour ajuster les paramètres d'extraction. Les valeurs extraites avec notre développement et celles extraites avec l'outil de référence étant très proches, nous avons tout de même appliqué un coefficient de correction de quelques pourcents à la plupart d'entre elles pour obtenir une précision encore plus fine.

Afin de valider ce développement, nous avons utilisé un circuit DES (Data Encryption Standard) Asynchrone développé et conçu au laboratoire TIMA [89] [57]. La validation finale a consisté à simuler le circuit électriquement. Nous avons utilisé dans un premier temps la "netlist" extraite du schéma, puis celle extraite du dessin des masques sans parasites, puis celle de l'extraction incluant les parasites extraits avec l'outil de référence StarRCXT et enfin celle du dessin des masques incluant les parasites extraits avec Diva. Le tableau 4.1 ci-dessous représente le temps de calcul du circuit et montre le résultat de ces simulations, faites avec le simulateur Nanosim.

schematic	**extract sans parasite**	**extract avec parasites StarRCXT**	**extract avec parasites Diva**
34 ns	**41 ns**	**114 ns**	**114 ns**

TAB. 4.1 – *Validation de l'extraction de parasites sous Diva*

On peut déduire de ce tableau que l'extraction des capacités parasites sous Diva donne des résultats de simulation identiques à ce que l'on obtient avec l'outil de référence et que par conséquent ce développement et cette implémentation sont conformes aux références. De plus ces résultats ont permis également de valider la méthode qui consiste à comparer des valeurs de composants parasites à partir de structures de tests de différentes formes.

En ce qui concerne l'extraction de parasites avec l'outil Assura qui utilise l'outil RCX, le développement de fichiers technologiques s'appuie sur les mêmes grandeurs physiques du procédé de fabrication, mais son implémentation est différente. Elle consiste à définir l'ensemble du procédé dans un fichier "procfile" où sont décrits tous les niveaux métalliques et diélectriques, leur résistivité et épaisseur, ainsi que quelques règles de dessins d'espacement et de taille minimum.

A partir de ce fichier et de 2 autres définissant les métaux devant être pris en compte entre autres, il est possible de générer un répertoire comportant un ensemble de fichiers binaires qu'utilise l'outil Assura / RCX lors de l'extraction. Ceci est réalisé grâce à un utilitaire "capgen" fourni dans Cadence /Assura. La méthode de validation a été la même que pour l'outil Diva, c'est à dire en utilisant des structures de tests. Ce développement permet donc la simulation post-layout précise de circuits hybrides

très complexes comportant les capacités parasites.

4.8 Simulation numérique "magnétique": description comportementale

Dans le cas de circuits numériques, comme nous l'avons décrit dans le chapitre dédié au flot de conception, les simulations ne sont pas électriques mais sont numériques. Afin de pouvoir simuler un circuit numérique intégrant des jonctions magnétiques, en utilisant la bascule compacte non volatile proposée dans nos travaux, il est nécessaire d'avoir une description comportementale de cette flip-flop, en Verilog. Cette description est ensuite compilée pour être interprétée par l'outil de simulation, Modelsim en l'occurrence. La description de cette flip-flop magnétique au format Verilog est décomposée en 5 modules, comme suit:

- gestion_com: ce module permet d'orienter les signaux "Ch1" et "Ch2" soit respectivement sur les MOS MP1 et MP2, soit sur respectivement MP2 et MP1, en fonction de la valeur de la sortie de la bascule. Sa description est une affectation par équation booléenne.
- MTJ: ce module permet de donner une valeur binaire aux jonctions tunnel. Il s'agit d'un module d'écriture des jonctions. Soit "0" pour une résistance faible dans le cas d'une orientation parallèle entre les 2 couches magnétiques, soit "1" pour une résistance forte dans le cas d'une aimantation antiparallèle. Les paramètres d'entrée sont Ch1 et Ch2 ainsi que la commande du champ d'écriture Hp et Hap. Le tableau 4.2 donne un exemple d'une partie de l'implémentation de ce module au format Verilog.

N°	Ch1	Ch2	Hp	Hap	Qt	Qt+1
1	0	0	0	1	?	0
2	0	0	1	0	?	1
3	0	0	0	0	?	-

TAB. 4.2 – *Implémentation en Verilog de l'état des jonctions tunnel magnétiques*

L'interprétation de ce tableau est la suivante:
N°1: Si Hp et Hap ont pour valeur respectivement '0' et '1', c'est à dire qu'un courant circule dans la ligne de champ d'écriture dans un sens, alors que Ch1 et Ch2 ont pour valeur '0', c'est à dire qu'un courant circule dans les jonctions et que par conséquent elles sont chauffées, alors la sortie prend '0', quelle que

soit sa valeur courante.

N°2: De la même façon, si Hp et Hap ont pour valeur respectivement '1' et '0', c'est à dire qu'un courant circule dans la ligne de champ d'écriture dans l'autre sens, alors que Ch1 et Ch2 ont pour valeur '0', c'est à dire qu'un courant circule dans les jonctions et que par conséquent elles sont chauffées, alors la sortie prend '1', quelle que soit sa valeur courante.

N°3: Si Hp et Hap ont pour valeur '0', c'est à dire qu'aucun courant ne circule dans la ligne de champ d'écriture, alors la sortie ne change pas, même si Ch1 et Ch2 ont pour valeur 0.

Toutes les combinaisons sont ainsi implémentées en définissant la table de vérité d'un registre, dans laquelle des notions de front peuvent figurer, "01" par exemple pour matérialiser un front montant sur un signal.

- latch_MAG: Ce module permet de définir la valeur de la sortie du latch magnétique en fonction de l'entrée D et du niveau haut ou bas de la clock, mais également en fonction de l'état des jonctions, et donc de Ch1, Ch2 et Az lors de la phase de lecture de la partie magnétique. Ce module est également décrit en utilisant la table de vérité d'un registre.
- latch_CMOS: Ce module défini classiquement le comportement d'un latch à partir de son entrée et du niveau haut ou bas de la clock.
- DFF_MAG: Il s'agit du module "top" de la hiérarchie, dans lequel sont instanciés tous les précédents modules et interconnectés entre eux, à l'image d'un "port map" en VHDL.

Cependant, bien que cette description n'apporte pas d'information sur les délais de propagation de la porte entre les entrées et les sorties, car dans un premier temps seule une simulation fonctionnelle est faite, ce type de description comportementale permet de donner un temps de propagation fixe d'une porte logique. Ce temps n'est donc pas du tout dépendant de l'environnement dans lequel se trouve la flip-flop. Cela permet tout de même de vérifier le fonctionnement d'un circuit sans considération d'optimisation de timing. De plus, il est important que les aspects de timing entre les signaux eux-mêmes soient définis et vérifiés lors de la simulation, car les contraintes sont précises en technologie TAS-MRAM, comme nous l'avons décrit sur le chronogramme de la figure 3.11. Par ailleurs, le format Verilog permet également de définir des contraintes de temps entre les signaux eux-mêmes. Soit des temps dit de "setup" soit de "hold" c'est à dire le temps de maintien d'un niveau sur un signal par rapport au changement de niveau d'un autre signal. Ces temps peuvent être définis entre 2 fronts montants, entre 2 fronts descendants ou entre 2 fronts différents.

La figure 4.7 illustre cet aspect de timing entre signaux.

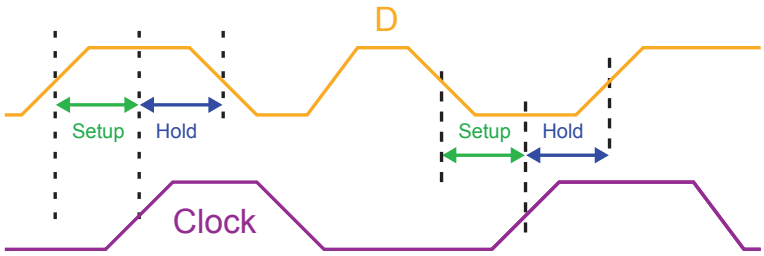


FIG. 4.7 – *Définition des temps de setup et de hold*

Si un temps de setup ou de hold n'est pas respecté dans les stimulis de simulation, alors l'outil de simulation Modelsim s'arrête et indique une erreur du type:
Error: *DFF_MAG.v(240): $width(posedge Hp:124 ns, :128 ns, 5 ns);*

Nous avons donc défini dans la description Verilog un ensemble de règle de timing relative à l'écriture des jonctions pour la technologie magnétique TAS-MRAM. Par exemple, il est indispensable que le signal Ch1 prenne une valeur '0' pendant 2 ns avant le signal Ch2. Ceci est donc défini de cette façon:
'define DFFMAG_Ch1_Ch2_negedge_negedge 2
$setuphold(negedge Ch1, negedge Ch2, 'DFF_MAG_Ch1_Ch2_negedge_negedge');

Ces règles définies dans le modèle de description comportementale permettent de respecter l'intégrité des signaux lors des simulations et de s'assurer que les phases de lectures et d'écriture respectent bien le mode de fonctionnement de la flip-flop non volatile que nous proposons. La figure 4.8 donne un exemple de simulation numérique sous Modelsim de la flip-flop non volatile compact que nous proposons dans ce kit de conception, basée sur la description Verilog présentée ci-dessus.

La simulation montre plusieurs phases de fonctionnement de cette bascule:

- **Phase 1:** Cette phase sert d'initialisation de la bascule. En effet, lors de la lecture la sortie de la bascule dépend de l'état des jonctions, et lors de l'écriture l'état des jonctions dépend de la sortie. Cette étape d'initialisation est donc indispensable. Dans le cas contraire, les signaux sont dans un état indéterminé.
- **Phase 2:** C'est une phase de lecture qui montre qu'après une lecture Q garde sa valeur "0" car les jonctions n'ont pas été réécrites.
- **Instant t1:** A cet instant, la bascule est en mode de fonctionnement CMOS classique. Q prend la valeur de D sur front montant de l'horloge, soit Q = 1.
- **Phase 3:** Il s'agit d'une phase de lecture qui montre que l'on restaure toujours la valeur stockée dans les jonctions. A la fin de cette phase, Q prend à nouveau la valeur '0'.

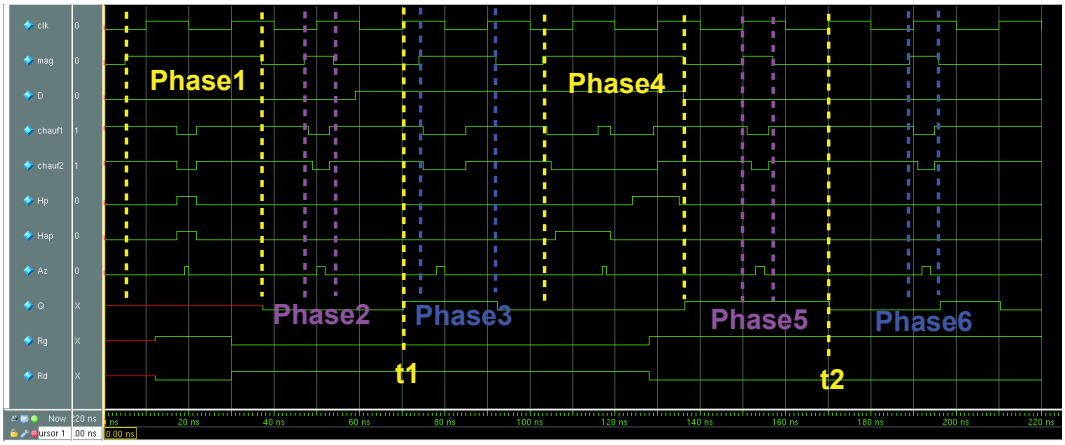


FIG. 4.8 – *Simulation numérique de la flip-flop non volatile compacte basée sur une description comportementale au format Verilog*

- **Phase 4:** Il s'agit d'une phase d'écriture pour laquelle on remarque que les signaux Rg et Rd représentant les résistances de la branche de gauche et de droite du latch magnétique prennent chacune une valeur opposée. A la fin de cette phase, Q prend également une valeur opposée à celle de la phase 1, soit Q = 1.
- **Phase 5:** C'est une phase de lecture qui montre qu'après une lecture Q garde sa valeur "1" car les jonctions n'ont pas été réécrites.
- **Instant t2:** La flip-flop est à nouveau dans un mode de fonctionnement CMOS classique. Q prend la valeur de D sur front montant de l'horloge, soit Q = 0.
- **Phase 6:** Il s'agit d'une phase de lecture qui montre que l'on restaure toujours la valeur stockée dans les jonctions. A la fin de cette phase, Q prend cette fois-ci la valeur '1'.

Cette figure 4.8 montre toutes les phases de fonctionnement de cette bascule, intégrant les mêmes spécificités que pour les simulations électriques. D'une part une écriture ou une lecture ne fonctionne que si le signal "mag" est actif (au niveau "1") afin d'isoler le latch magnétique du reste de la flip-flop. D'autre part, la mise à jour de la sortie de la bascule peut être soit synchrone avec le signal "mag" si celui-ci est désactivé sur niveau haut de l'horloge, soit la sortie peut être synchrone avec l'horloge si le signal "mag" est désactivé sur niveau bas de l'horloge. Cette description Verilog intègre également les aspects de timing entre les signaux "magnétiques" pour l'écriture et la lecture. En revanche, il est difficile d'intégrer des notions de séquencement notamment pour l'écriture. Ceci implique que la valeur des résistances Rg et Rd ne change pas successivement comme pour une simulation électrique mais

simultanément, en même temps que la seconde résistance. Néanmoins, à la fin d'une phase d'écriture, les 2 jonctions sont systématiquement dans un état opposé, ce qui est le fonctionnement souhaité et réel.

4.9 Simulation numérique "magnétique": timing et rétro annotation

Dans un deuxième temps, les simulations numériques doivent intégrer les notions réalistes de timing des cellules, afin d'obtenir un résultat proche du comportement du circuit fabriqué. Il est donc nécessaire de rétro annoter les simulations en injectant dans l'outil un fichier de timing SDF, pour Standard Delay Format. Pour cela, l'outil utilise un fichier .lib pour chaque bibliothèque utilisée, dans lequel tous les aspects de timing sont définis. Nous avons donc caractérisé électriquement la flip-flop afin d'extraire toutes ces informations. Cette étape consiste à simuler cette porte logique avec le modèle compact pour la simulation électrique, afin de déterminer les délais de propagation de la cellule.

Le délai dans une porte est dépendant de plusieurs paramètres. D'une part de la complexité de la cellule et du dimensionnement de ses transistors, et d'autre part de son environnement extérieur. En effet, le temps de propagation dépend de la rapidité du signal d'entrée, appelé slope, ainsi que de la capacité de charge sur la sortie. Pour chacun de ces paramètres est définie une table à plusieurs valeurs, appelé Look Up Table, ce qui permet d'obtenir l'ensemble des combinaisons entre toutes les valeurs de slope et toutes les valeurs de capacités. Plus ces tables sont riches, plus la simulation numérique sera précise. Chaque table comporte généralement 5 valeurs, ce qui permet d'avoir 25 cas de configurations différentes. Ci-dessous un exemple de LUT:

lu_table (table_27)
variable_1 : input_net_transition ;
variable_2 : total_output_net_capacitance ;
index_1 (" 0.0048, 0.1088, 0.2608, 0.5248, 1 ");
index_2 (" 0.0057, 0.0137, 0.0297, 0.0817, 0.1617 ");

Le comportement temporel d'une porte logique est décrit et défini selon 2 critères. D'une part le temps de propagation entre le signal d'entrée et le signal de sortie, et d'autre part le temps de montée et de descente du signal de sortie. Les temps de setup et de hold décrits précédemment sont donnés pour Vdd / 2, alors que les temps de propagation sont donnés entre la tension Vtn du signal d'entrée et la tension Vtp du signal de sortie, et que les temps de montée et de descente de la sortie sont donnés

de 10% à 90%. Il y a donc dans le fichier .lib 4 sections pour le signal Q de la flip-flop, chacune comportant les 25 configurations, et pour 3 cas de variation: minimum, typique et maximum. Nous avons donc extraits tous ces paramètres par simulation électrique pour définir le .lib de la flip-flop non volatile que nous proposons. Cela permet de générer un fichier SDF pour l'ensemble du circuit, c'est à dire pour toutes les cellules qu'elles soient CMOS classique ou CMOS / Magnétique. Nous avons effectivement vérifié lors de la simulation que les temps de propagation des portes logiques n'est alors plus de 0.1ns qui est une valeur par défaut pour chacune d'entre elles, mais ceux implémentés dans le fichier de description de timing.

4.10 Placement et Routage "magnétique"

Afin de pouvoir utiliser la flip-flop non volatile dans le flot de placement et de routage standard en utilisant également des cellules standard de bibliothèque, son dessin des masques nécessite de respecter certaines règles de géométrie, dépendant de la technologie, STMicroelectronics 130n dans le cas du développement de ce kit de conception. Les principales règles que nous avons suivies pour faire le layout sont les suivantes:

- chaque cellule doit avoir un rail Vdd en haut et un rail Gnd en bas.
- La largeur des rails d'alimentation est fixe.
- La hauteur d'une cellule est fixe. Sa largeur par contre peut être variable mais doit être un multiple d'un pas de grille fixe.
- Les pins d'entrée et de sortie doivent se trouver centrées sur le pas de grille.
- Chaque cellule doit respecter toutes les règles DRC. Il en est de même après aboutement à n'importe quelle autre cellule de bibliothèque.
- L'encombrement, X et Y, de chaque cellule doit être défini par un masque spécifique identique à toutes les cellules.

Depuis la vue "layout", nous avons pu créer une vue "abstract" à partir d'outil utilitaire, "abstract gen" en l'occurrence. Cette vue est utilisée pour la génération du fichier LEF (Library Exchange Format) dans lequel sont définis entre autres tous les aspects de géométrie de la cellule, pour tous les rails, toutes les connexions métal et VIAx, toutes les pins d'entrée / sortie. De façon générale, pour optimiser cette étape de placement et de routage, l'outil peut retourner les cellules de 180°. Le rail Gnd est alors en haut et le rail Vdd en bas, une rangée sur deux. L'outil peut d'autre part faire un miroir selon l'axe des X ou des Y pour positionner les pins stratégiquement afin de simplifier le routage. En revanche, dans le cas de notre cellule magnétique, il

est important qu'elle ne soit ni retournée ni utilisée en miroir selon Y, car le sens du courant d'écriture générant le champ doit être maitrisé et connu, identique à toutes les flip-flops. En effet, si une cellule est retournée de 180° alors le champ appliqué lors d'une écriture sera l'opposé de celui escompté. Ces contraintes impliquent que les flip-flops magnétiques ne soient placées qu'une rangée sur deux sur le layout.

Le procédé hybride CMOS / Magnétique mis en place par le consortium lors du projet ANR impose que les lignes de champ soient sur le dernier niveau de métal CMOS, soit le métal5. Afin de respecter cette contrainte de fabrication, le dessin des masques de la flip-flop comporte un rail de metal5 le long des jonctions tunnel. Cependant, pour que ces rails soient connectés ensemble et aux générateurs de courant d'écriture, il est indispensable d'ajouter dans le flot de placement et routage magnétique une étape spécifique. Cette étape s'insère entre la phase de placement et la phase de routage. En effet, lorsque toutes les cellules du circuit sont placées, il est nécessaire de positionner des pistes de metal 5 au-dessus de chacune des rangées de cellules comportant des flip-flops magnétiques. Afin de placer celles-ci exactement en superposition avec le layout des cellules, il faut indiquer à l'outil de placement et routage le nombre de rails à placer, ainsi qu'un offset en haut et en bas pour indiquer la première et la dernière rangée. La figure 4.9 donne un exemple pour lequel les lignes de champ ont été placées une rangée sur deux au-dessus des flip-flops magnétiques, pour un circuit où toutes les cellules ont été préalablement placées.

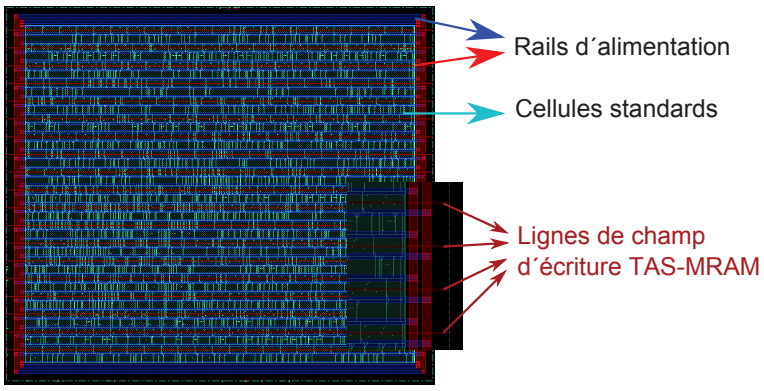


FIG. 4.9 – *Placement automatique de cellules standard et des rails de ligne de champ d'écriture pour technologie TAS-MRAM*

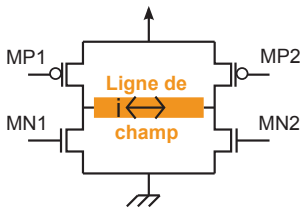
Une autre étape spécifique à cette technologie hybride est celle du routage de l'arbre d'horloge. En effet, contrairement à un circuit CMOS synchrone standard qui n'a la plupart du temps qu'un seul signal d'horloge, la flip-flop que nous proposons utilise 2 signaux clk1 et clk2. Ceci permet d'isoler le latch magnétique pendant une

phase de lecture ou d'écriture des jonctions. Dans ce flot spécifique, il est donc nécessaire de router ces 2 signaux en premier avant l'ensemble des signaux du circuit. Le signal clk2 est tout de même prioritaire car c'est sur front montant de clk2 que la sortie de la bascule prend la valeur de l'entrée. La fin du flot de placement et routage reste le même que celui pour un circuit CMOS classique.

4.11 Générateur de courant pour l'écriture TAS

4.11.1 Implémentation et architecture

Le courant d'écriture permettant de générer le champ de retournement des jonctions est fourni par un générateur de courant. Afin de pouvoir coder les deux états magnétiques dans les jonctions ce courant doit être bidirectionnel. Pour cela nous utilisons comme présenté sur la figure 4.10 deux générateurs de part et d'autre de la ligne de champ.



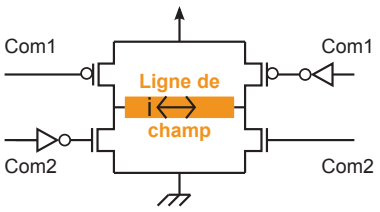
i: courant d´écriture		MP1 - MP2			
		00	01	10	11
MN1 MN2	00	0	0	0	0
	01	cc	>0	cc	0
	10	cc	cc	<0	0
	11	cc	cc	cc	0

FIG. 4.10 – *Générateur de courant d'écriture des jonctions tunnel magnétiques*

On peut remarquer sur cette figure que le fait d'avoir une commande propre à chaque transistor impose la gestion de 4 signaux et par conséquent le besoin d'avoir 4 plots d'entrées. D'autre part, la table de vérité montre que sur les 16 combinaisons, seules 2 permettent de générer un courant soit dans un sens soit dans l'autre, 7 autres ne permettent à aucun courant de circuler et 7 autres créent un court-circuit entre Vdd et Gnd.

Nous avons donc étudié plusieurs configurations de commande des signaux MP1, MP2, MN1 et MN2 à partir de 2 signaux de commande seulement. Chacune d'entre elles présente ses avantages et ses inconvénients. La figure 4.11 montre par exemple que sur les 4 configurations possibles du tableau de Karnaugh, 2 permettent de générer le courant soit dans un sens soit dans l'autre, par contre les 2 autres configurations créent un court-circuit entre Vdd et Gnd à l'état de repos, ce qui est catastrophique du point de vue consommation, échauffement, vieillissement etc.

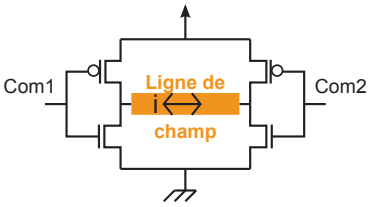
Parmi toutes les configurations étudiées, la configuration optimum que nous proposons est celle présentée sur la figure 4.12 pour laquelle il y a toujours 2 com-



i: courant d´écriture		Com1	
		0	1
Com2	0	cc	<0
	1	>0	cc

FIG. 4.11 – *Générateur de courant d'écriture des jonctions tunnel magnétiques à 2 signaux de commandes*

binaisons "10" et "01" qui permettent de générer le champ soit dans un sens soit dans l'autre, et deux combinaisons correspondantes à l'état de repos "00" et "11", c'est à dire lorsque que l'on n'est pas dans une phase d'écriture. Ces 4 combinaisons sont très conventionnelles et permettent de simplifier la génération de ces signaux de commandes.



i: courant d´écriture		Com1	
		0	1
Com2	0	0	>0
	1	<0	0

FIG. 4.12 – *Générateur de courant d'écriture des jonctions tunnel magnétiques à 2 signaux de commandes optimisés*

L'aspect architectural des générateurs de courant est important mais leur dimensionnement l'est tout autant. C'est l'objet de l'étude du paragraphe suivant.

4.11.2 Dimensionnement

Afin de générer suffisamment de champ pour le retournement de l'aimantation de la couche de référence des jonctions tunnel, il est nécessaire de dimensionner correctement les générateurs de courant. Le champ nécessaire est imposé par le post-process magnétique. Dans le cadre du projet ANR SPIN pour lequel ce kit de conception a été développé, les spécifications du champ étaient 30 Oersted. La valeur du champ magnétique H (en A/m) généré par une ligne de champ de longueur "l", de largeur "w", d'épaisseur "t", dont le centre se trouve à une distance "z" du centre de la couche de stockage avec un facteur de "cladding" (figure 4.13), et parcourue par un courant I k_{clad}, est donnée par l'*equation 4.3*:

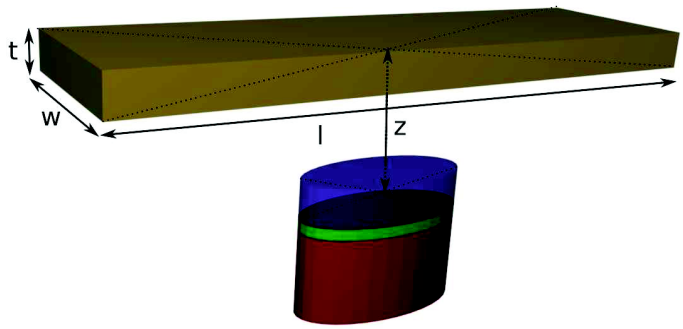


FIG. 4.13 – *Conventions utilisées pour le calcul du champ magnétique*

$$H\left(I,k_{clad},z,l,w,t\right) = BS\left(k_{clad},z,l,w,t\right) \times I \tag{4.3}$$

avec $BS\left(k_{clad},z,l,w,t\right)$ un facteur calculé à partir de la loi de Biot et Savart qui donne le champ généré en un point par une ligne de champ élémentaire et aprés intégration sur tout le volume de la ligne de champ. Ce coefficient est donné par l'*equation 4.4*:

$$BS\left(k_{clad},z,l,w,t\right) = \frac{k_{clad}}{4 \times \pi \times w \times t} G\left(z,l,w,t\right) \tag{4.4}$$

avec :

$$\begin{aligned} G\left(z,l,w,t\right) \quad = \quad & \left[F\left(-\frac{l}{2},-\frac{w}{2},-z-\frac{t}{2}\right) - F\left(\frac{l}{2},-\frac{w}{2},-z-\frac{t}{2}\right)\right] - \\ & \left[F\left(-\frac{l}{2},+\frac{w}{2},-z-\frac{t}{2}\right) - F\left(\frac{l}{2},\frac{w}{2},-z-\frac{t}{2}\right)\right] - \\ & \left[F\left(-\frac{l}{2},-\frac{w}{2},-z+\frac{t}{2}\right) - F\left(\frac{l}{2},-\frac{w}{2},-z+\frac{t}{2}\right)\right] - \\ & \left[F\left(-\frac{l}{2},\frac{w}{2},-z+\frac{t}{2}\right) - F\left(\frac{l}{2},\frac{w}{2},-z+\frac{t}{2}\right)\right] \end{aligned} \tag{4.5}$$

et :

$$F\left(a,b,c\right) = \frac{b}{2}\log\frac{k\left(a,b,c\right)+a}{k\left(a,b,c\right)-a} + \frac{a}{2}\log\frac{k\left(a,b,c\right)+b}{k\left(a,b,c\right)-b} - c\arctan\frac{ab}{ck\left(a,b,c\right)} \tag{4.6}$$

et :

$$k\left(a,b,c\right) = \sqrt{a^2+b^2+c^2} \tag{4.7}$$

Pour obtenir la valeur du champ H en A/m, il faut multiplier le champ magnétique en Oe par 79,57. Le coefficient BS est dépendant des géométries et donc

du procédé CMOS et post-process magnétique. Dans le cas de ce projet ANR, le champ de 50 Oe correspond à un courant théorique de 14 mA circulant dans la ligne d'écriture placée sous les jonctions.

Cependant, cette ligne métallique est plus ou moins résistive en fonction de sa longueur et de sa largeur. Plus elle est longue plus elle est résistive, et plus elle est large moins elle l'est. En ce qui concerne le courant électrique circulant dans la ligne de champ, il est également fonction de la résistance de cette ligne, car plus la ligne est résistive, plus la chute de tension est importante, donc plus les tensions Vds des transistors du générateur sont faibles. Par conséquent, le courant diminue de la même façon. C'est ce qu'illustre de façon schématique la figure 4.14.

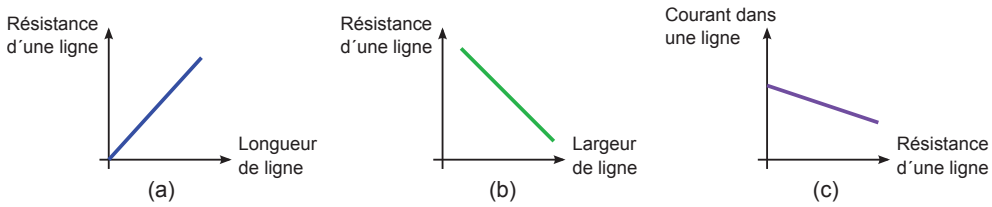


FIG. 4.14 – *Evolution de la résistance d'une piste métallique en fonction de sa longueur (a) et de sa largeur (b), Evolution du courant dans une ligne en fonction de sa résistance (c).*

La longueur de la ligne est plus ou moins imposée par la complexité du circuit et du nombre de cellules. Plus le circuit est complexe plus sa taille est importante et plus la longueur des lignes de champ augmente. Cependant, il est possible d'imposer une taille rectangulaire pour le coeur du circuit afin de réduire sa largeur et donc la longueur des lignes de champ. Cela implique que la hauteur du coeur du circuit augmente ainsi que le nombre de rangées de flip-flops. Le nombre de lignes d'écriture augmente en conséquence, ce qui n'est pas un inconvénient. Pour le dimensionnement des lignes de champ, la tendance serait alors d'augmenter fortement la largeur des lignes pour diminuer sa résistance, et donc augmenter le courant circulant sous les jonctions. Cependant, le modèle de simulation TAS des jonctions permet de montrer que plus la largeur de la piste augmente, plus le courant nécessaire pour générer le champ de retournement est important, ce qui va à l'encontre du raisonnement précédent. Ceci implique qu'il y a un compromis à trouver entre la largeur des lignes de champ et le dimensionnement des transistors du générateur de courant pour une longueur de ligne donnée, connue après l'étape du floorplan du placement et routage. Afin de trouver le bon compromis, nous avons caractérisé l'évolution du courant en fonction de la résistance de la ligne pour plusieurs longueurs de ligne ainsi que le

courant nécessaire en fonction de cette résistance. Dans le cas d'une fonderie dédiée à ce type de fabrication, la distance entre la ligne de champ et la jonction pourrait être plus faible, ce qui permettrait d'utiliser des courants plus faibles également.

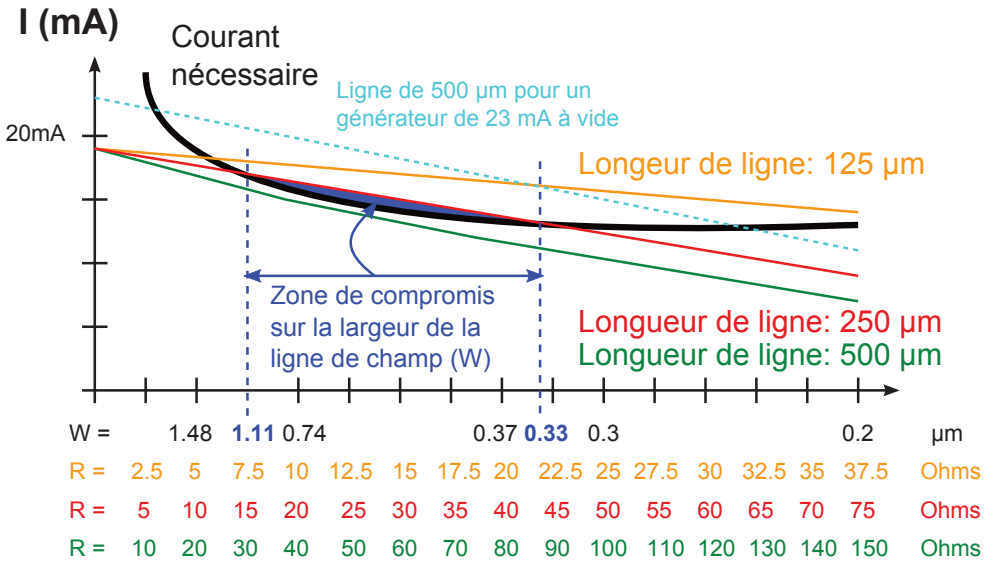


FIG. 4.15 – *Dimensionnement de la largeur de la ligne de champ*

L'illustration de la figure 4.15 montre clairement la zone dans laquelle le compromis doit se faire en fonction de la longueur de la ligne. Pour la technologie ST 130n sur laquelle nous avons travaillé, la résistance typique est de 60 mOhm/carré. Prenons l'exemple de la courbe rouge correspondant à une ligne de champ de 250 μm de longueur. W étant la largeur de la ligne de champ, si W>1.11 μm ou si W<0.33 μm, alors le courant circulant dans la ligne de champ est inférieur au courant nécessaire au retournement, représenté par la courbe noire. Dans ces 2 configurations il n'est pas possible d'écrire les jonctions. En revanche, la zone bleue montre que pour cette même ligne de 250 μm de long, si 1.11 μm < W < 0.33 μm alors le champ généré sera suffisant pour retourner les jonctions. On remarque aussi sur cette figure que pour une ligne d'une longueur de 500 μm représentée par la courbe verte, un générateur dimensionné pour 19 mA sans résistance de charge ne permet en aucun cas d'écrire une jonction car cette ligne verte est systématiquement en dessous la courbe noire, matérialisant le courant nécessaire au retournement d'une jonction. Dans ce cas, il suffit de dimensionner un générateur capable de fournir 23 mA à vide par exemple, cas de la courbe bleu ciel pointillée, au lieu de 19mA. Les dimensions des transistors sont les suivantes: $(W/L)_{Pmos} = 70\mu m/0.13\mu m$ et $(W/L)_{Nmos} = 34\mu m/0.13\mu m$. Enfin on peut déduire de cette illustration que plus le circuit sera complexe plus la ligne sera longue, donc plus la piste sera large sans toutefois dépasser une limite

d'environ 1.1 μm, et sans nécessairement être obligé de concevoir un générateur de courant bien plus important. Nous pouvons donc conclure de cette étude que le dimensionnement des générateurs de courant doit être précis et qu'il est dépendant de la géométrie finale du circuit.

Hormis les aspects de champ nécessaire au retournement, nous avons également étudié les phénomènes d'électro migration dans ces lignes de champ dans lesquelles un fort courant circule. L'électro migration est un phénomène de déplacement d'atomes dans un conducteur induit par un flux d'électrons. Ce mécanisme n'apparait que dans les applications où l'on observe de très fortes densités de courant [51]. Ce phénomène a été découvert en 1861 par le physicien français Gerardin [40].

La relation permettant d'évaluer le courant maximum autorisé dans une ligne en fonction de sa taille est la suivante:

$$I = I_{eqMax}(Wmin) + (W - Wmin) \times I_{eqMax}(W = 1\mu m)/1\mu m \tag{4.8}$$

avec:

$I_{eqMax}(Wmin)$: 1.41 mA @ 70°C (donnée du DRM STMicroelectonics)

W : *largeur de la ligne*

$Wmin = K_L \times Ieq \times f(T) + \Delta w$ (largeur minimum de la ligne de champ)

où: $K_L = \frac{1}{Epaisseur\ minimum\ Cuivre \times Jmax}$

et: $Epaisseur\ minimum\ Cuivre = 325\ nm$

et: $Jmax = 3.72\ mA\ /\ \mu m^2$

où: $Ieq = \frac{\int_0^{Tc} i(t)\,\mathrm{d}t}{Tc}$

Le cas le plus défavorable pour Ieq est 15 mA DC, dans un mode de fonctionnement où les données sont sauvegardées à chaque cycle d'horloge.

où: $f(T) = e^{[\frac{Ea}{nKbT0}(1-\frac{T0}{T})]}$

et: $Ea = 0.8\ eV$; $n = 2$; $Kb = 1.38.10^{-23}$; $1eV = 1.6.10^{-19}$; $T = 300K$; $T = 398K$

où: $\Delta W = 0.02\ \mu m$ (donnée du DRM STMicroelectronics)

On obtient donc Wmin = $0.28\mu m$ soit Imax = 7.05 mA @ 70°C.

Considérons le compromis étudié précédemment où la largeur de la ligne de champ est $W = 1\mu m$: dans ce cas le courant Ieq = 7.5 mA soit environ Imax. En revanche si les sauvegardes sont ponctuelles, 1 cycle sur 10 par exemple, le courant $Ieq = 1.25mA << Imax$.

Notons que ces spécifications d'électro migration sont données pour avoir moins de 0.1% de fautes en 10 ans. Nous pouvons donc considérer que les règles d'électro migration sont respectées dans notre dimensionnement. De plus, sachant que l'écriture des jonctions se fait en 2 phases et que pour chacune de ces phases le courant circule dans un sens opposé, le courant IeqMax est en théorie nul pour un courant bidirectionnel d'après la relation $Ieq = \frac{\int_0^{Tc} i(t)\,\mathrm{d}t}{Tc}$.

4.11.3 Dessin des masques en vue du placement-routage et insertion dans le flot de conception

L'étude que nous avons faite sur le dimensionnement des générateurs de courant est valable pour un circuit full custom ainsi que pour un circuit numérique. L'objectif que nous nous sommes fixé est d'intégrer le placement et le routage de ces générateurs dans le flot de conception numérique. Lors de la phase de placement, l'outil offre la possibilité de placer 1 colonne d'une cellule sur la gauche et une autre sur la droite du coeur du circuit. Dans un circuit CMOS classique, ces cellules sont appelées ENDCAP et servent à éviter que des zones actives de cellules standards soient placées en bordure de circuit. Nous avons utilisé cette possibilité pour placer les générateurs de courant. Cela implique donc d'avoir une version "gauche" et une version "droite" des composants. De plus, nous avons fait 2 versions de chaque, l'une permettant de générer 20mA l'autre permettant de générer pour 25 mA.

En ce qui concerne le routage de ces générateurs, il ne peut être fait lors de la phase de routage car les cellules ENDCAP ne sont pas considérées comme cellules fonctionnelles. Nous avons donc fait en sorte qu'après aboutement des générateurs aux cellules du coeur, la connexion soit automatique. Ceci implique que les générateurs fournissent le courant sur le niveau metal5. Ainsi, les rails de metal5 servant aux lignes de champ se juxtaposent avec celles en metal5 des générateurs. De plus,

les cellules flip-flops ne pouvant être placées qu'une rangée sur deux, nous avons optimisé l'utilisation des générateurs en faisant en sorte que tous ceux qui sont placés sur des rangées où il n'y a pas de flip-flop magnétique soient tout de même utilisés. De même, les commandes des grilles sont dessinées de façon à ce qu'elles soient toutes connectées par aboutement. La figure 4.16 montre comment une telle conception du dessin des masques des générateurs de courant permet de simplifier et d'automatiser leur routage. La seule connexion qui doit être faite manuellement est celle entre les plots "Com1" et "Com2" vers les deux grilles des générateurs.

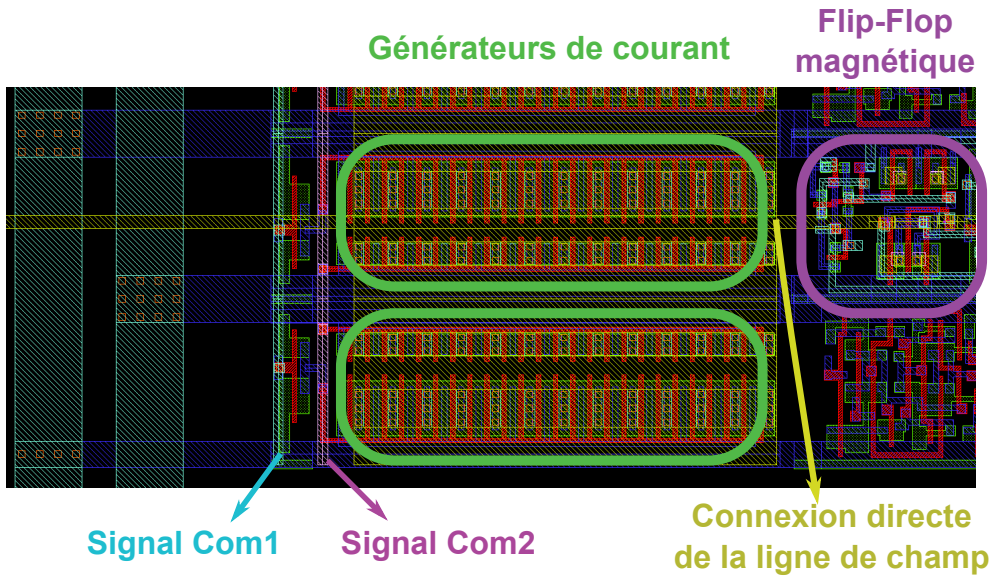


FIG. 4.16 – *Placement et routage automatique des générateurs de courant d'écriture*

4.12 Conclusion

Lors de projets de recherche, il est parfois difficile de prouver des résultats théoriques sur des idées ou des concepts. En microélectronique, la preuve sur silicium est très reconnue et appréciée par la communauté, c'est pourquoi un des objectifs des projets ANR CILOMAG et SPIN était de concevoir et fabriquer des échantillons de dispositifs hybrides CMOS / magnétiques innovants. Le développement de ce kit de conception et la mise en place de flots de conception spécifiques a permis à différents partenaires de concevoir de tels circuits à partir d'outils logiciels industriels performants auxquels ils étaient habitués. Ce kit et ces flots permettent de couvrir toutes les étapes de conception à la fois d'un circuit full custom et numériques. Cela permet donc de concevoir, dessiner, vérifier, simuler précisément des circuits beaucoup plus complexes que les premiers démonstrateurs qui ont pu être réalisés auparavant.

C'est l'objet du prochain chapitre dans lequel deux types de circuits numériques sont présentés, selon deux types d'application que nous avons identifié pour lesquels un procédé CMOS/Magnétique présente un intérêt, ainsi qu'une étude comparative de consommation.

Chapitre 5

Intégration de Jonctions Tunnel Magnétiques dans un circuit intégré complexe

5.1 Introduction

Toute application électronique a ses propres contraintes, étroitement liées aux besoins de l'environnement dans lequel elle se trouve. De nos jours, la consommation des dispositifs est un paramètre important. C'est pourquoi nous avons consacré une partie de ce chapitre à une étude de consommation, comparant les performances en termes d'énergie statique d'un circuit CMOS standard et celle d'un circuit hybride CMOS / Magnétique, pour un système donné. Cette étude est établie à partir d'une application à base d'un processeur simple. Les codes source sont ceux utilisés en enseignement. Ils nous ont été fournis par un enseignant. Cette étude permet d'illustrer un concept nouveau. Par ailleurs, la non volatilité permet d'apporter d'autres avantages à des circuits microélectroniques, notamment la fiabilité et la sécurité. C'est l'objet de la première partie de ce chapitre dans lequel nous proposons la conception d'un circuit numérique haute sécurité, à l'aide du kit de conception que nous avons présenté dans le chapitre précédent.

5.2 Application haute sécurité: filtre numérique non volatile

5.2.1 Description

L'idéal pour une application haute sécurité serait de mémoriser en permanence l'état du circuit, dans les registres de bas niveau non volatils, afin de pouvoir restaurer un état stable à tout moment. Le concept est illustré sur la figure 5.1 qui est une simulation électrique d'un compteur / décompteur, pour lequel toutes les valeurs des bascules sont sauvegardées à chaque cycle d'horloge dans leurs propres jonctions tunnel. Les signaux représentés sur cette illustration sont l'alimentation, l'horloge, le signal "MAG" actif pour chaque opération de lecture ou d'écriture des jonctions, les sorties DATA<0:3>, ainsi que l'état magnétique des 2 jonctions MTJ1 et MTJ2 pour chaque bit de sortie.

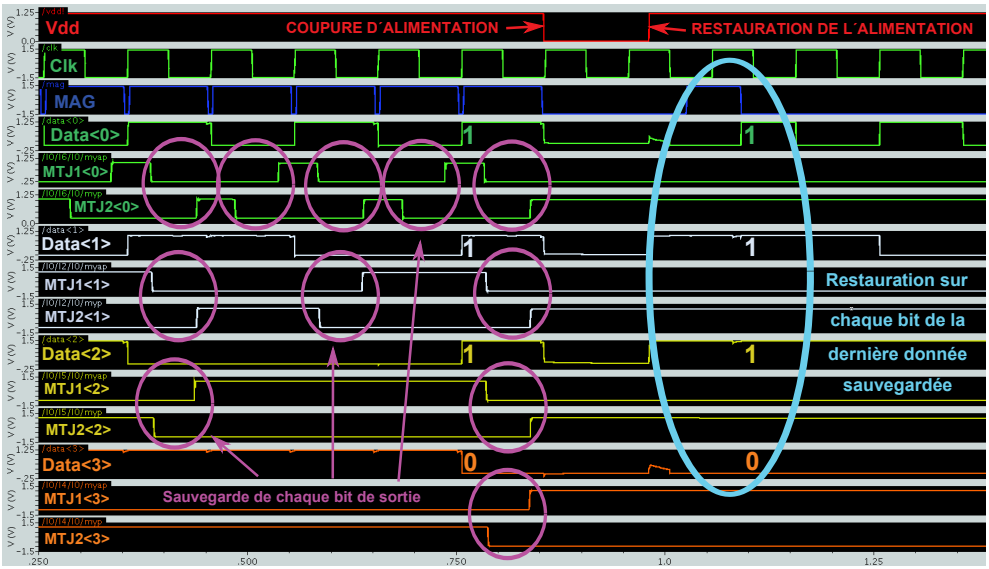


FIG. 5.1 – *Restauration de données après une coupure d'alimentation*

Nous pouvons noter qu'à chaque cycle d'horloge l'état magnétique des 2 jonctions change systématiquement lorsque le niveau de la sortie correspondante a été modifié. La dernière valeur sauvegardée sur Data<0:3> est "1110", puis survient une coupure d'alimentation où tous les niveaux sont alors à '0'. A la restauration de l'alimentation, les bascules prennent une valeur aléatoire, "0110" sur Data<0:3>. On observe ensuite une phase de restauration du dernier état sauvegardé, où l'on peut alors remarquer que le Data<0:3> prend de nouveau la valeur "1110". Cette coupure d'alimentation n'a donc pas engendré de dysfonctionnement ni de remise à zéro du compteur /

décompteur.

Pour illustrer cette application haute sécurité sur un circuit numérique plus complexe qu'un simple compteur, l'exemple que nous avons choisi est un filtre numérique à réponse impulsionnelle finie de 32 coefficients, qui a pour objectif de filtrer les fréquences basses. Ce circuit reçoit sur son entrée *Filter_In* la valeur du signal échantillonné par un convertisseur analogique / numérique et délivre en sortie un signal filtré à un convertisseur numérique / analogique. Le schéma de principe est présenté sur la figure 5.2

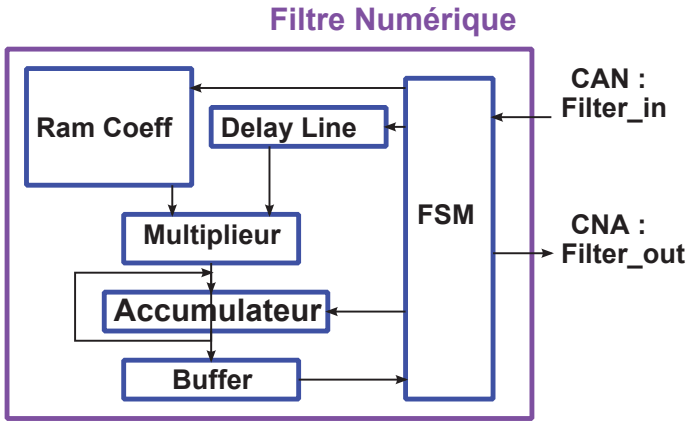


FIG. 5.2 – *Schéma de principe du filtre FIR*

Ce filtre comporte plusieurs parties comme suit:

- une machine à état permettant de contrôler le séquencement des calculs et la génération de signaux vers le CAN et le CNA.
- un module de registre à décalage "delayline" permettant de charger les échantillons fournis par le CAN à la fin de la conversion seulement.
- un module "coeff_ram" permettant de stocker les 32 coefficients du filtre pas bas, préalablement calculé à partir du logiciel Matblab.
- un module "multiplieur" assurant les multiplications successives des 32 coefficients par les échantillons.
- un module "accu" qui permet de faire la somme de toutes les multiplications pour un même échantillon.
- un module "buffer" qui permet de synchroniser le résultat en sortie à la fin du calcul pour ne fournir que le bon résultat et non pas des valeurs intermédiaires incohérentes.

Le fonctionnement de ce filtre est le suivant: l'entrée est multipliée successivement par les 32 coefficients, puis une somme globale est faite par additions successives. Si

le résultat montre que la fréquence est inférieure à la fréquence de coupure alors elle est transmise en sortie, si elle est supérieure elle est filtrée et la sortie vaut '0'. Il se passe donc plusieurs dizaines de cycles d'horloge entre 2 échantillons disponibles en sortie du filtre. Le résultat de simulation de ce filtre est présenté sur la figure 5.3.

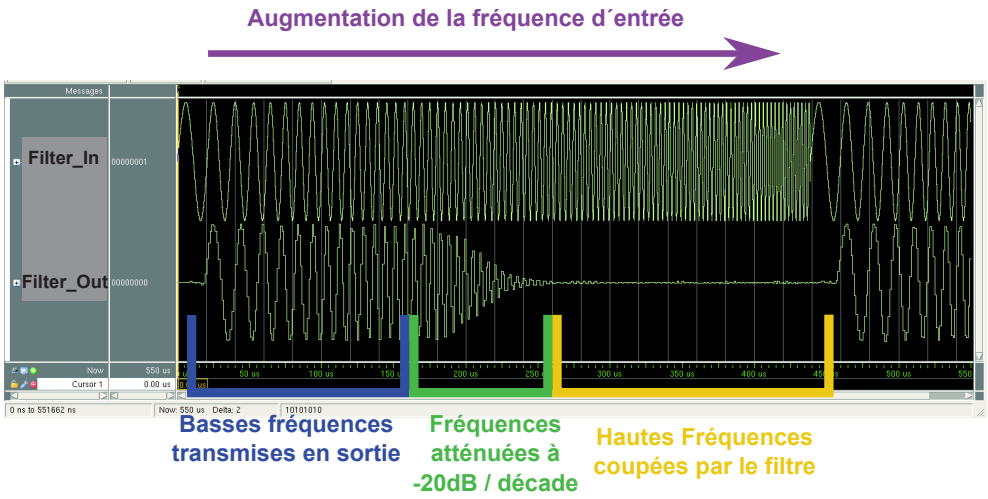


FIG. 5.3 – *Simulation du filtre FIR*

Une façon d'apporter de la fiabilité à ce circuit est de sauvegarder tous les registres pendant les phases de calcul, à tous les cycles. C'est ce que permet de faire l'intégration de jonctions tunnel dans le circuit, en remplaçant toutes les flip-flops par des flip-flops magnétiques. En effet, il est dans ce cas possible de sauvegarder à chaque cycle d'horloge le contenu de chacune des bascules dans sa propre partie magnétique. Ceci permet de sécuriser les données de calcul, car s'il y a une coupure de courant intempestive pendant la phase de calcul il est toujours possible de revenir dans le dernier état stable et de poursuivre le calcul de filtrage sans devoir recommencer avec un autre échantillon. De même, il est possible de sauvegarder l'information de sortie du filtre de façon non volatile pendant toute la durée de traitement de chaque échantillon et fournir au DAC une valeur cohérente à n'importe quel moment.

5.2.2 Implémentation sur technologie CMOS / Magnétique

A partir du kit de conception et du flot que nous avons développé, nous avons transformé ce filtre numérique CMOS en filtre numérique CMOS / Magnétique non volatil. Les différentes étapes de la conception sont la synthèse logique, la simulation numérique, le placement et routage pour terminer avec les vérifications LVS. Nous présentons ci-après chacune de ces étapes et décrivons les spécificités d'utilisation de bibliothèque hybride.

5.2.2.1 Synthèse logique sur technologie CMOS / Magnétique

Les sources nous ayant été fournies, au format VHDL, nous avons dû mettre en place une stratégie de modification pour intégrer la bascule non volatile que nous proposons. En effet, cette flip-flop innovante a des signaux spécifiques à la commande de la partie magnétique. La liste de tous les signaux est donc la suivante: D et Q pour les signaux classiques, Az, Ch1 et Ch2 pour les signaux propres aux JTM. La gestion des signaux Hp et Hap, servant à la génération du champ d'écriture, est identique à celle présentée au cours du chapitre dédié au kit de conception. La connexion de ces 2 terminaux se fait automatiquement lors de la phase de placement / routage. Enfin, rappelons que cette cellule n'a pas un seul signal d'horloge mais 2, clk1 et clk2, générés automatiquement par un module spécifique à partir des signaux clk et mag.

Dans la description des composants du fichier .lib propre à chaque bibliothèque, il ne semble pas possible de définir des cellules avec des ports inutiles. S'ils ne sont pas utilisés, ils sont par défaut connectés au potentiel Gnd, ce qui serait faux et surtout un problème pour notre circuit. Or les signaux dits "magnétiques" ne peuvent pas prendre de sens du point de vue de la synthèse. Cela implique que ces signaux ne sont pas inclus dans la description de la bascule pour la synthèse du fichier .lib. Cependant ils sont présent physiquement. Dans ce cas, il suffit d'ajouter ces signaux magnétiques à chaque entité des modules séquentiels, ainsi qu'un signal clock commun. L'entité "seq_CMOS" devient alors l'entité "seq_mag" de la façon suivante:

```
port(                                          port(
In_A  : in  std_logic;                         In_A  : in  std_logic;
In_B  : in  std_logic;                         In_B  : in  std_logic;
clk   : in  std_logic;   devient               clk   : in  std_logic;
Out   : out std_logic;                         clk2  : in  std_logic;
port);                                         Az    : in  std_logic;
                                               Ch1   : in  std_logic;
                                               Ch2   : in  std_logic;
                                               Out   : out std_logic;
                                               );
```

TAB. 5.1 – *Modification d'une entité VHDL pour la synthèse logique "magnétique"*

Ainsi la synthèse peut alors être faite en utilisant uniquement des flip-flops magnétiques. La liste des noeuds après synthèse a donc tous les ports magnétiques sur les modules séquentiels. Il suffit d'appliquer un script très simple pour ajouter les

ports magnétiques Az, Ch1, Ch2 et Clk2 à la liste des ports des bascules et par la même occasion de leur attribuer un noeud. Ce script s'exécute en une seule ligne de la façon suivante:

sed s/.clk2(clk)/.Az(Az), .ch1(ch1), .ch2(ch2), .clk1(clk1), .clk2(clk)/ ASIC_ cmos.v > ASIC_ mag.v

On remplace le port "clk2" connecté au noeud "clk" de toutes les bascules par la liste des ports magnétiques chacun connecté sur un noeud commun, soit le port Az sur le noeud Az, le port "ch1" sur le noeud "ch1" etc. De cette façon, toutes les bascules sont connectées aux ports correspondants de l'entité à laquelle elles appartiennent. Ce "port map" logiciel est présenté sur la figure 5.4 sur laquelle les connexions bleues et noires sont faites de façon classique au moment de la synthèse et les connexions vertes sont faites par application de la commande "sed" présentées ci-dessus. Cette "netlist" permet de passer aux étapes suivantes de la conception, c'est à dire la simulation numérique et le placement / routage pour concevoir le dessin des masques en vue de la fabrication.

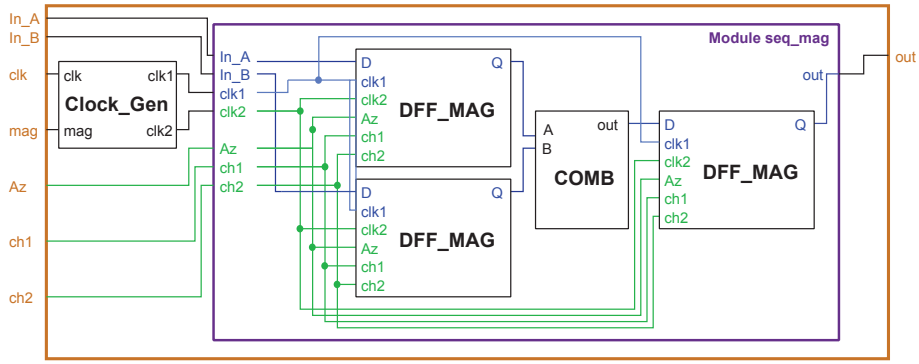


FIG. 5.4 – *Schéma de principe des connexions aux bascules non volatiles*

5.2.2.2 Simulation d'un filtre numérique sur technologie CMOS / Magnétique

Pendant cette phase de conception du filtre, nous avons souhaité valider le fonctionnement global de notre bascule innovante et l'implémentation de son comportement au format Verilog. Nous avons donc fait le même type de simulation que pour la version CMOS, à savoir appliquer en entrée une sinusoïde à fréquence variable pour observer le filtrage à partir d'une certaine fréquence. Le résultat de simulation a montré que notre bascule se comporte comme attendu et que la fonction de filtrage est assurée de la même façon. Ensuite nous avons simulé la partie magnétique. La figure 5.5 montre premièrement une phase de sauvegarde à l'instant t1 = 91 μs. Lors

de cette phase, toutes les flip-flops du circuit sont sauvegardées en même temps, dans leur propre partie magnétique. Le filtre continue tout de même de fonctionner normalement. Deuxièmement à l'instant t2 = 101 μs, on observe une phase de récupération de l'état de l'ensemble du circuit, en restaurant les données contenues dans les jonctions tunnel vers les sorties des bascules. On remarque alors que la sortie du filtre prend une valeur inattendue, qui est celle de l'instant t1, correspondant à la dernière sauvegarde.

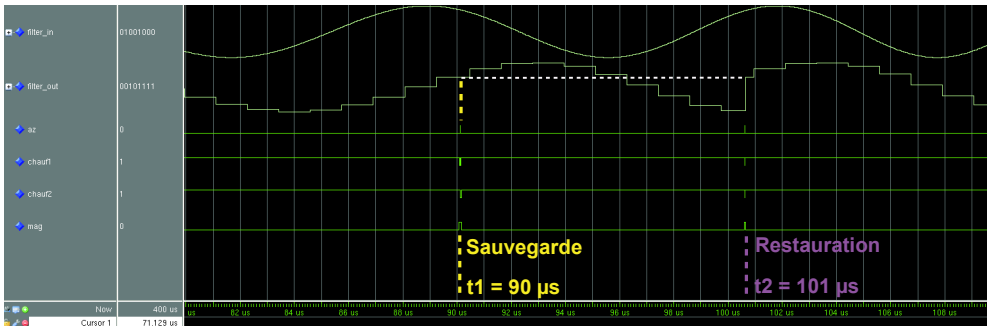


FIG. 5.5 – *Simulation d'un filtre numérique non volatil sur technologie CMOS / Magnétique*

Cette simulation valide d'une part l'implémentation du comportement de notre bascule au format Verilog, et montre d'autre part que le flot et le kit de conception que nous proposons permettent de concevoir des circuits numériques non volatils à partir des outils de conception industriels standards. Le fait d'être capable de synthétiser et de simuler tout un circuit avec des bascules non volatiles peut entre-autre être intéressant pour des applications haute sécurité dans lesquelles chaque état peut être sauvegardé à chaque coup d'horloge. Cela permet de restaurer un état connu et stable à n'importe quel moment.

5.2.2.3 Placement et routage d'un filtre numérique sur technologie CMOS / Magnétique

La méthode que nous avons suivie pour cette étape est celle décrite dans le chapitre dédié au kit de conception que nous proposons. Le rapport de synthèse indique que ce circuit comporte 1070 cellules standard de bibliothèques dont 292 bascules non volatiles, ainsi que 1637 cellules de remplissage "filler". La surface du coeur du circuit est de 230 μm x 210 μm. Le résultat de l'étape de placement et routage est présenté sur la figure 5.6.

La dernière étape de conception de ce filtre est la vérification LVS. Nous avons

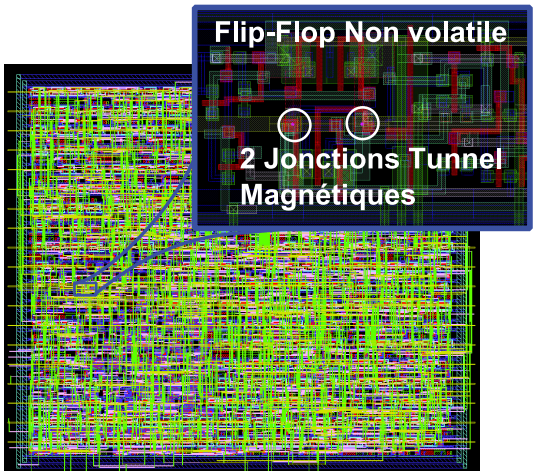


FIG. 5.6 – *Placement et routage d'un filtre numérique non volatil*

donc importé le fichier de description du circuit après synthèse faite sous Synopsys / Design Vision au format Verilog sous l'environnement d'édition de schéma Cadence / Schematic Composer, puis le fichier de description extrait de la vue layout, placé et routé avec le logiciel Cadence / Velocity, sous l'environnement d'édition de dessin des masques Cadence / Virtuoso. Le LVS a été fait à partir du kit de conception magnétique présenté précédemment. Le résultat de cette vérification a été concluant, le dessin des masques est bien conforme au schéma.

5.3 Etude de consommation d'un circuit intégré

La deuxième partie de ce chapitre est consacrée à une étude de consommation des circuits microélectroniques qui est aujourd'hui une des plus fortes contraintes des applications industrielles. Que ce soit pour l'autonomie des appareils portables, pour le vieillissement des batteries ou pour l'environnement, réduire la consommation d'un circuit intégré est primordial. Il existe plusieurs méthodes pour réduire la consommation, notamment la consommation statique. En effet, comme le montre la figure 5.7, la consommation à l'état de repos prend une part très importante dans les technologies très avancées. Cette évolution a été très rapide à partir des procédés inférieur à 90nm [22], [19].

Les tailles des transistors sont de plus en plus petites donc les densités d'intégration augmentent fortement, les MOS sont de plus en plus rapide donc les épaisseurs d'oxyde sont de plus en plus fine, ce qui accroît considérablement les courants de fuite et donc la consommation dite de standby. Nous présentons dans ce chapitre quelques techniques qui permettent de réduire cette consommation statique et dynamique,

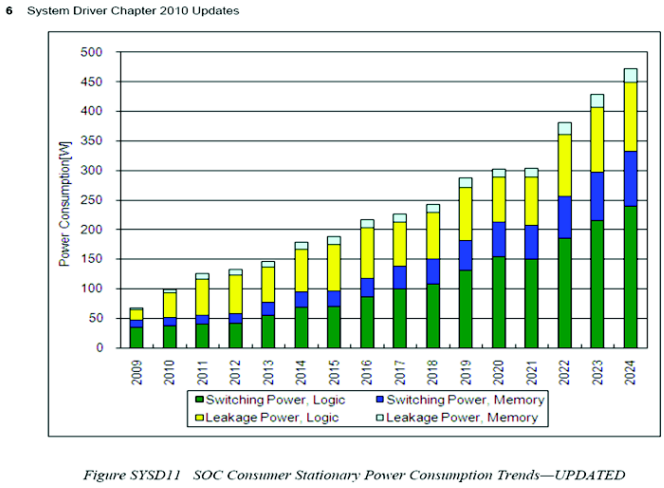


FIG. 5.7 – *Evolution de la consommation dans les circuits intégrés*

et notamment grâce à l'utilisation de jonctions tunnel magnétiques qui rendent les circuits CMOS non volatils.

5.3.1 Techniques de conception pour la faible consommation

L'objectif de cette section sur la faible consommation des circuits intégrés n'est pas de comparer toutes les techniques ni de prendre position, mais plutôt de faire un état de l'art des techniques les plus répandues à ce jour, pour ensuite situer nos travaux et notre méthode parmi celles existantes. Les systèmes complexes utilisent des techniques de gestion de l'énergie au niveau système en faisant de la gestion dynamique de la puissance [15] [27], en s'appuyant sur des stratégies de fonctionnement d'un circuit complexe. Cela consiste à déterminer les périodes pour lesquelles le circuit est en état de repos seulement ou de sommeil, si cette période est plus ou moins longue afin de déterminer si la partie du circuit concernée peut être mis dans un état de repos ou de sommeil sans que cela ne coûte de l'énergie plutôt que d'en gagner [80]. Dans les paragraphes suivants, nous présentons seulement des techniques de conception bas niveau, c'est à dire plus au niveau circuit que système, en abordant les thèmes des substrats SOI, les techniques de variation ou de coupure de l'alimentation et de l'horloge, totalement ou partiellement. Il s'agit des techniques de DVFS, clock gating et power gating.

5.3.1.1 Substrat SOI

L'industrie de la microélectronique utilise divers types de matériaux semi-conducteurs sous forme monocristalline tel que le silicium, le plus utilisé, mais aussi des composés

d'éléments des familles III/V et II/VI du tableau périodique [123]. Dans la grande majorité des applications, seule la partie superficielle des plaquettes de matériau semi-conducteur est utile à l'implantation de la partie active du circuit. La plus grande partie de l'épaisseur sert en effet essentiellement de support mécanique. Deux technologies utilisant le silicium peuvent être distinguées: la technologie bulk (figure 5.8 (a)) et la technologie SOI (Silicon On Insulator), silicium sur isolant (figure 5.8 (b)), pour laquelle l'épaisseur du silicium est d'une cinquantaine de nanomètre seulement.

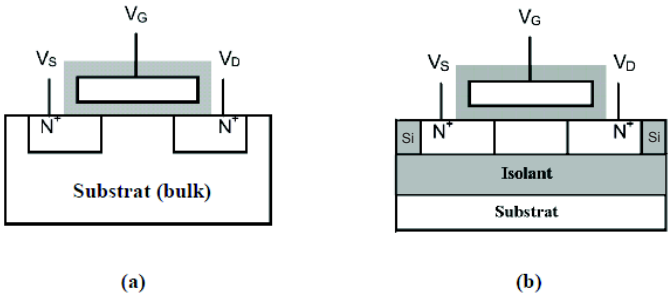


FIG. 5.8 – *Vue de coupe de transistors NMOS en technologie Si-bulk (a) et SOI (b)*

Dans le cas d'un transistor MOS conventionnel en technologie Si-bulk ou massif, on peut constater que les zones dopées qui sont les zones actives, sont directement implantées dans une masse de silicium épaisse, le substrat. La présence d'un substrat épais en continuité électrique avec les couches superficielles induit des phénomènes parasites dans celles-ci et les rend sensibles notamment à des perturbations électriques, notamment des courants de fuite vers le substrat. Ces courants de fuite apparaissent essentiellement au niveau des jonctions drain/source et substrat, ou drain/source et caisson, ainsi que caisson et substrat. Ils provoquent donc une augmentation de la puissance consommée par le circuit. La technologie SOI a permis de minimiser, voire supprimer, grâce à l'isolation des parties actives par rapport au substrat, ce phénomène de courant de fuite, entre autres [63]. L'inconvénient de cette technologie est principalement son coût plus élevé, ce qui est de moins en moins vrai avec l'avancement des noeuds technologiques. L'utilisation de substrats SOI reste encore marginal de nos jours mais pourrait prendre une part plus importante à partir des technologies 45nm et 32nm [62]. La part de marché des wafers SOI était de 2% en 2004 et de 11% en 2009 [73].

5.3.1.2 Méthode DVFS: Dynamic Voltage and Frequency Scaling

Cette méthode DVFS pour "Dynamic and Voltage Frequency Scaling" consiste à réduire soit la tension, soit la fréquence de certaines parties d'un circuit intégré en fonction de leur utilisation. La plupart du temps, les deux paramètres sont ajustés simultanément tant ils sont liés du point de vue des performances. Lorsqu'un bloc est inactif, sa tension d'alimentation peut être réduite, et ainsi diminuer les courants de fuite. Cela peut également être le cas lorsque ce même bloc ne requiert pas une rapidité de calcul, car le fait de diminuer sa tension d'alimentation augmente le temps de commutation des transistors. Dans ce cas, la réduction de la fréquence a le même effet sur la vitesse et en plus sur la consommation dynamique. L'adaptation de la fréquence permet également de réduire la température du circuit. Un inconvénient de cette méthode est d'une part de fournir au circuit une tension variable ou plusieurs tensions fixes, et d'autre part de gérer l'alimentation de plusieurs parties de circuit indépendamment, par logiciel ou matériel selon les cas. De fait, l'intégration de cette gestion augmente la taille du circuit, donc un surcoût, car il est nécessaire d'intégrer un bloc de gestion, comme illustré sur la figure 5.9. Il a été montré que cette augmentation est de 12% sur une application multi processeurs [28]. Enfin, ces deux techniques peuvent parfois avoir un impact sur les performances d'un circuit, selon le mode d'implémentation.

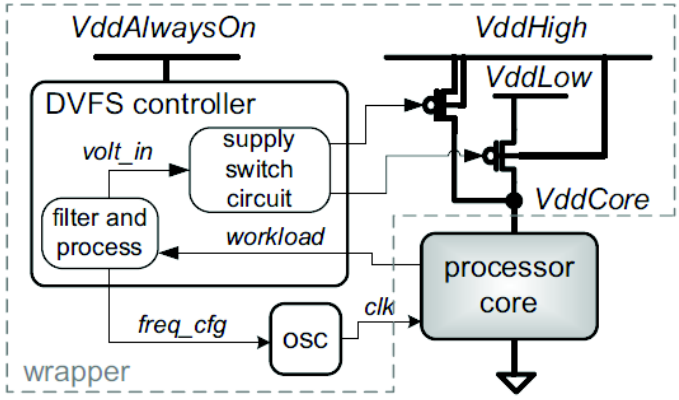


FIG. 5.9 – *Schéma de principe de la technique DVFS*

Ces techniques sont particulièrement implantées dans les applications de type ordinateurs portables ou autres appareils mobiles fonctionnant sur batterie, pour lesquels la consommation est un paramètre très critique.

5.3.1.3 MOS en fonctionnement sous le seuil

Le principe d'un circuit en fonctionnement sous le seuil est de réduire fortement la tension de grille, jusqu'à ce qu'elle soit inférieure au Vt des transistors. Ceci peut s'appliquer soit au circuit, soit à une partie seulement. Cette forte réduction de la tension peut être considérée selon différents modes de fonctionnement. Premièrement lorsque les circuits n'ont besoin que ponctuellement de hautes performances [25], deuxièmement lorsque la contrainte principale est la consommation et que le circuit ne requiert pas de très hautes performances. C'est le cas de circuits micro capteurs ou implants. Et troisièmement, lorsque certaines parties sont ponctuellement inactives. Cela peut être le cas de mémoire SRAM par exemple [117]. Les inconvénients de cette méthode sont la faible transconductance (courant de polarisation faible), un fonctionnement lent et une forte sensibilité à l'appariement des tensions de seuil (mismatch) [54].

5.3.1.4 Clock gating

Cette méthode consiste à ne pas propager l'horloge jusqu'à certaines parties d'un circuit d'un système complexe. Cela permet de réduire la consommation dynamique globale car en effet, un module séquentiel qui reçoit un signal d'horloge pendant une phase où il n'est pas sollicité dans le fonctionnement, consomme énormément d'énergie. Cette méthode n'a aucune influence sur la consommation statique. Du point de vue de la conception il est possible lors de la synthèse logique de choisir si l'on souhaite intégrer une notion de clock gating, module par module. Dans ce cas, l'outil de synthèse fait un calcul booléen et ajoute un latch muni d'une entrée "enable" qui autorise ou non le signal d'horloge à se propager dans les différents blocs, couplé à une porte NAND, comme illustré sur la figure 5.10.

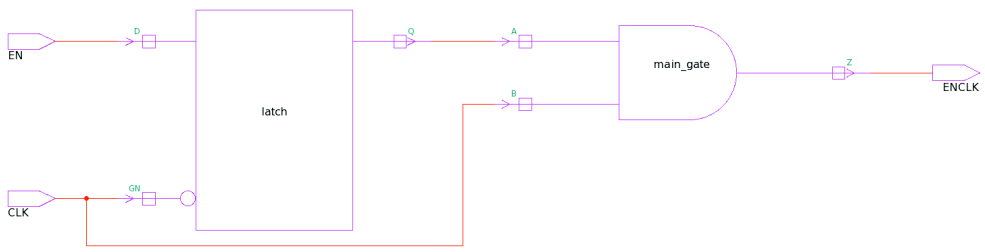


FIG. 5.10 – *Synthèse logique avec "clock_gating"*

5.3.1.5 Power gating

Cette méthode consiste à couper l'alimentation des parties inactives d'un circuit, ayant pour but essentiellement la réduction de la consommation statique. Cette méthode est basée d'une part sur le mode de fonctionnement du circuit mais également sur une stratégie de gestion de l'alimentation. Il a été montré qu'en utilisant plusieurs modes de veille différents, le gain sur l'énergie statique pouvait être jusqu'à 17% supplémentaires [10]. La coupure d'alimentation de certaines parties du circuit se fait soit entre Vdd et le circuit, soit entre le circuit et Gnd. On parle alors de "header switch" ou de "footer switch" (figure 5.11).

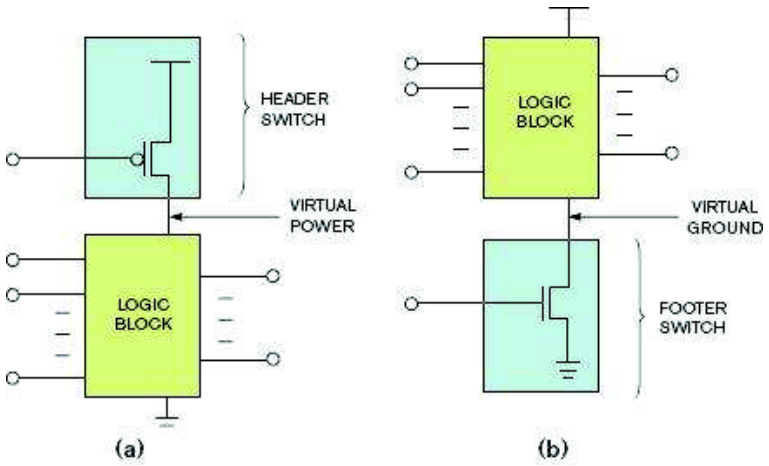


FIG. 5.11 – *Principe du power gating*

Néanmoins, le choix et le dimensionnement des transistors de coupure est un élément clef. Cette étape peut être faite par un concepteur de cellules standards mais des modules peuvent également être fournis sous forme d'IP. Par ailleurs cette méthode souffre d'un inconvénient qui est le redémarrage d'un module inactif, car les transistors utilisés sont très lent [10]. Cette méthode affecte dans la plupart des cas l'architecture des circuits, et un compromis est à trouver entre le gain en consommation et la remise en activité des modules inactifs. La mise en veille peut se faire soit par logiciel, soit par matériel. Synopsys a proposé une méthode de conception basée sur leur outils et une technologie 90n [3], dans lequel les aspects de contrôle stratégique des courants de fuite et de méthode de conception sont abordés.

5.3.2 CMOS versus CMOS/Magnétique

Depuis le début des mémoires MRAMs, la non volatilité a été l'atout principal des circuits hybrides CMOS/magnétiques. Dans nos travaux, nous nous sommes inté-

ressés aux circuits intégrés et plus particulièrement à la logique non volatile. Comme nous l'avons présenté précédemment, le remplacement des bascules CMOS par des bascules non volatiles permet de viser des applications haute sécurité et fiabilité d'information. Cependant, nous avons souhaité nous intéresser également aux applications faible consommation. En effet, il est instinctif de dire que le fait de pouvoir sauvegarder un état électrique dans des composants magnétiques et de pouvoir ensuite restaurer ces valeurs simplifie les différentes techniques présentées ci-avant, à savoir le clock gating et le power gating. Cependant, la phase de sauvegarde requiert une quantité d'énergie qui peut s'avérer non négligeable. Rappelons que pour une technologie TAS, le courant de chauffage est de l'ordre de quelques centaines de μA pendant quelques dizaines de ns. Il n'est donc pas raisonnable d'imaginer sauvegarder l'ensemble des bascules dans une application "low power". La courbe de la figure 5.12 montre la différence de consommation à l'état de repos entre un circuit CMOS et un circuit hybride.

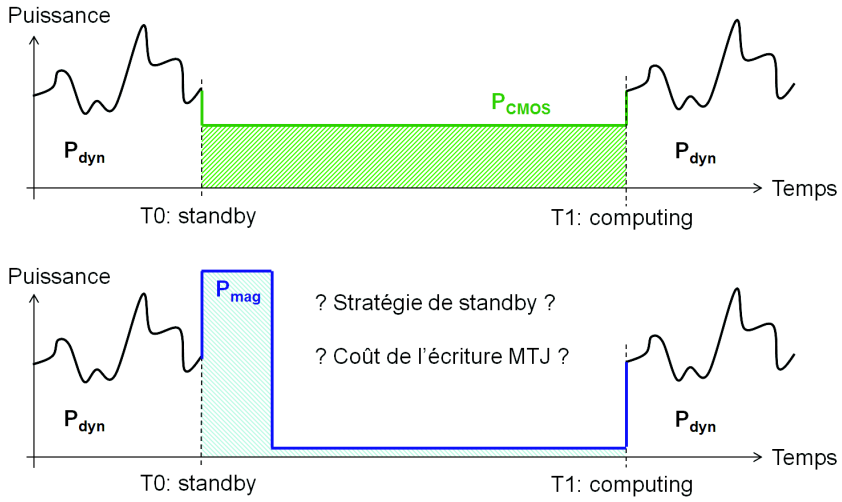


FIG. 5.12 – *Consommation d'un ASIC CMOS vs CMOS / Magnétique*

Lorsque le circuit fonctionne et effectue des calculs, alors la puissance dynamique consommée est la même que ce soit pour un circuit CMOS que pour un circuit CMOS/magnétique. En revanche lorsque le circuit est inactif, dans le cas d'un circuit CMOS, la puissance consommée est constante et est linéaire dans le temps, alors que pour un circuit hybride, la puissance consommée est d'abord élevée pendant une courte durée, ce qui correspond à la phase de sauvegarde, puis quasi nulle ensuite pendant toute la phase d'inactivité.

Il est donc nécessaire de modéliser l'énergie consommée dans ces 2 cas afin de pouvoir les comparer. Pour un circuit CMOS, l'énergie est dépendante du procédé

CMOS et plus particulièrement des courants de fuites I_{off} des transistors. Elle peut se modéliser selon l'*equation 5.1*.

$$E_{CMOS} = P \times t = U \times I \times t = V_{dd} \times I_{off} \times t_{stdby} \quad (5.1)$$

Pour un circuit hybride, l'énergie est aussi dépendante du procédé CMOS, car le dimensionnement des transistors permettant de générer le courant de chauffage I_{on} en dépend, mais également de l'application en fonction du nombre de registres à sauvegarder, ainsi que du procédé magnétique qui impose le courant et le temps nécessaire pour le chauffage et l'écriture. Cette énergie peut se modéliser selon l'*equation 5.2*.

$$E_{MAG-TAS} = E_{chauf.} + E_{retourn.}$$

$$E_{MAG-TAS} = [\, V_{dd} \times I_{on} \times Nbr_{FF_MAG} \times (t_{chauf}) \,] + [V_{dd} \times I_{champ} \times t_{champ} \times Nbr_{generateurs} \,] \quad (5.2)$$

Les 2 courbes de la figure 5.13 représentent l'évolution de la consommation statique d'un circuit CMOS, linéaire en fonction du temps, et celle d'un circuit hybride, quasi nulle après la phase de sauvegarde. On s'aperçoit qu'il y a un compromis à trouver entre le temps d'inactivité du circuit et le nombre de bascules à sauvegarder. Dès lors que le temps de standby est supérieur à t1 alors l'énergie consommée par la sauvegarde des registres est amortie, et il devient intéressant du point de vue de la consommation d'utiliser les jonctions tunnel magnétiques pour sauvegarder les données.

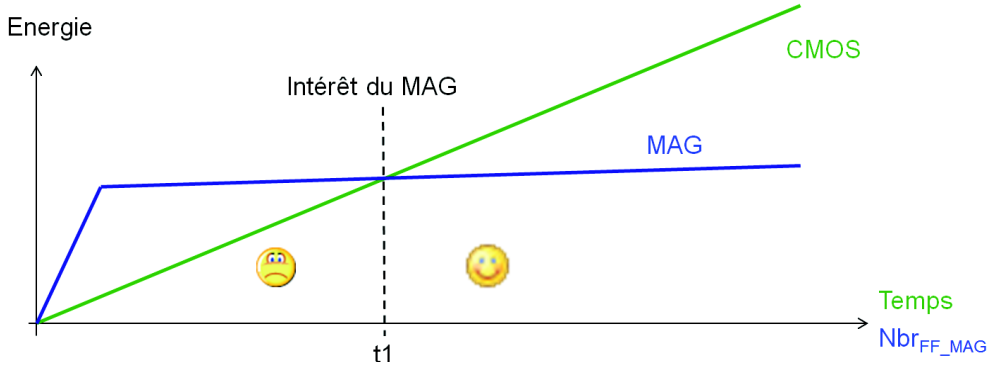


FIG. 5.13 – *Seuil d'énergie pour l'utilisation de jonctions tunnel magnétiques en vue de la réduction de la consommation statique*

Cette figure montre clairement que le choix et l'intérêt d'un tel procédé hybride CMOS/magnétique dépend essentiellement de l'application. C'est pourquoi nous proposons une étude de consommation selon plusieurs noeuds technologiques, sur une

application de type processeur simple pour la démodulation d'un signal, décrit dans la suite de ce manuscrit.

Dans cette étude nous avons considéré d'une part le procédé magnétique TAS, mais également le procédé STT pour lequel l'énergie consommée lors de l'écriture est définie par l'*equation 5.3*. On remarque qu'elle se résume à une seule composante. En effet, la méthode d'écriture STT ne requiert pas de champ extérieur car l'écriture se fait par courant polarisé et par conséquent l'énergie nécessaire est moindre par rapport au procédé TAS. De plus, rappelons qu'avec la méthode d'écriture TAS, une des deux composantes ne diminue pas avec le noeud technologique.

$$E_{MAG-STT} = V_{dd} \times I_{on} \times Nbr_{FF_{MAG}} \times t_{ecr} \tag{5.3}$$

5.3.2.1 Cas d'étude: processeur simple

Afin de faire cette étude de consommation, nous avons choisi une application d'un système complet. Ce système reçoit en entrée un signal modulé en fréquence, de type FSK (Frequency Shift Keying), puis filtré par un filtre actif analogique. En sortie de ce dernier, le signal est modulé d'une part en fréquence mais également en amplitude, puis il est numérisé par un convertisseur analogique / numérique. Le processeur, qui est l'élément de notre étude, reçoit ce signal et a pour objectif de le démoduler. La figure 5.14 montre la chaine de démodulation ainsi que la forme du signal à chaque étape.

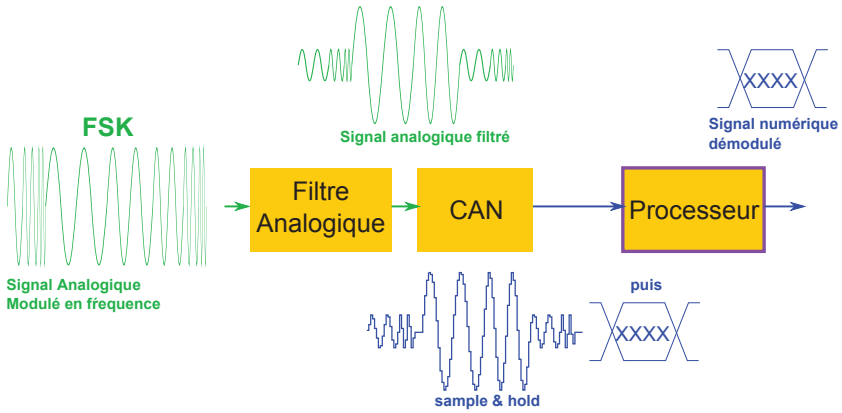


FIG. 5.14 – *Chaine de démodulation*

L'application que nous avons choisie requiert une haute définition, c'est pourquoi nous avons besoin d'un convertisseur analogique / numérique 16 bits. Classiquement, les types de convertisseurs capables de fournir une telle résolution sont des convertisseurs de type compteur ou double rampe, chacun fonctionnant à la même

fréquence, typiquement 100 KHz [64]. Dans cette étude, nous nous sommes intéressé au processeur seulement dans le but de mesurer le gain éventuel en consommation en utilisant un procédé CMOS/magnétique plutôt qu'un procédé CMOS standard. L'architecture de ce processeur est présenté sur la figure 5.15.

Ce processeur comporte 16 registres de données, une unité arithmétique et logique munie d'un registre accumulateur, un registre de résultat et d'un bus d'adressage. Le tout est cadencé par un séquenceur qui permet l'exécution du programme.

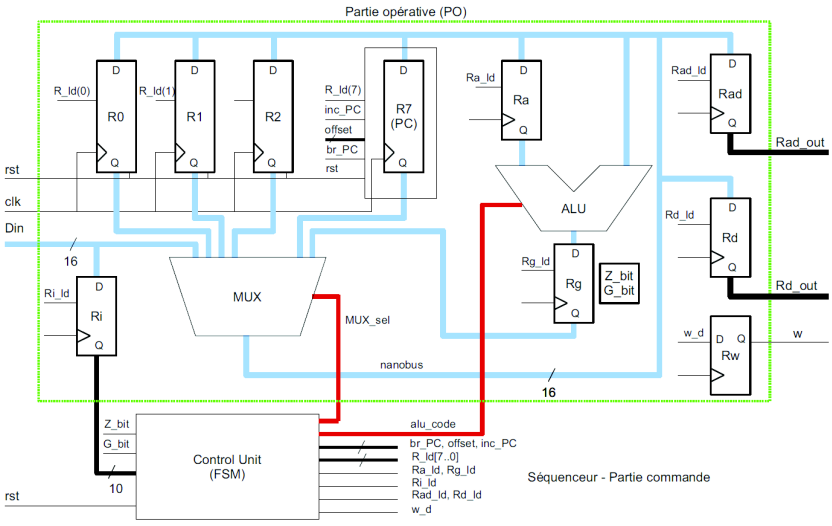


FIG. 5.15 – *Architecture du processeur simple cas d'étude*

A une fréquence de 100 KHz, ce convertisseur analogique / numérique est capable de fournir un mot de 16 bits toutes les 10 μs au processeur, qui lui fonctionne à une fréquence de 100 MHz. Nous avons déterminé par simulations numériques que le temps de traitement du processeur requiert au maximum 70 cycles d'horloge, soit 700 ns. Dans la mesure où le temps de calcul du processeur dépend du nombre de cycles d'horloge, la fréquence peut être adaptée. Elle doit être suffisamment élevée pour que le calcul soit terminé avant qu'un nouvel échantillon ne soit disponible. La vitesse de fonctionnement global de ce système, illustré sur la figure 5.16, est donc imposée par le convertisseur CAN.

En effet, entre l'instant où le processeur fourni en sortie la donnée, et l'instant où il reçoit une nouvelle donnée à traiter, s'écoule un certain temps d'inactivité du processeur. Dans ce contexte, notre étude a pour but d'évaluer s'il y a un éventuel gain de consommation, selon plusieurs technologies CMOS, et de quantifier ce gain.

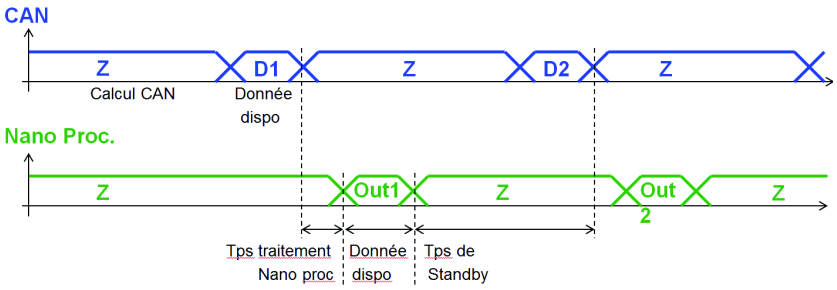


FIG. 5.16 – *Fonctionnement temporel du système*

5.3.2.2 Processeur magnétique: description

Afin de pouvoir faire cette étude, nous avons synthétisé ce processeur selon plusieurs technologies CMOS afin d'établir des comparaisons. Pour cela nous avons dû adapter les codes sources du processeur pour le rendre non volatil partiellement. En effet, comme nous l'avons décrit précédemment, il n'est pas utile de sauvegarder l'ensemble des registres du circuit et ce serait de plus beaucoup trop gourmand en énergie consommée. Dans l'application que nous présentons, seule la donnée de sortie doit pouvoir être restituée au module suivant en cas de besoin. L'architecture que nous proposons ici, illustrée sur la figure 5.17, est d'ajouter un registre de 16 bits en sortie, composé de 16 bascules non volatiles. Cela permet soit de couper l'alimentation de tout le processeur pendant toute la phase d'inactivité, soit de couper l'alimentation de la partie CMOS uniquement afin de garder l'information de sortie active. De plus, cela permet de restaurer instantanément la dernière valeur stable en cas de coupure d'alimentation intempestive de l'ensemble du système.

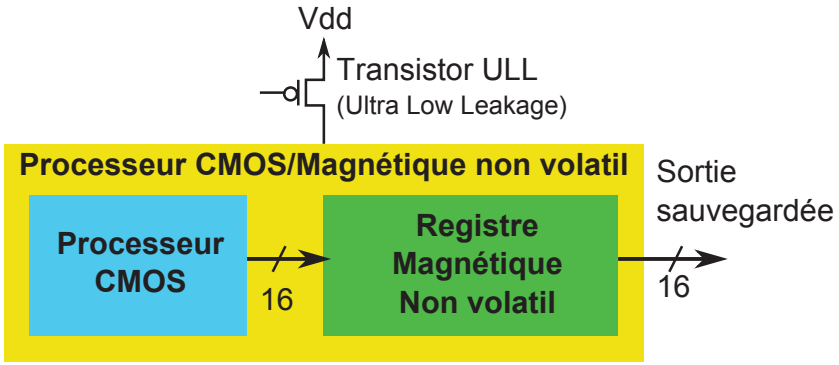


FIG. 5.17 – *Architecture du processeur magnétique non volatil*

On remarque sur cette figure que l'architecture intègre un transistor de coupure d'alimentation, de façon schématique car il en faut en réalité beaucoup plus en fonc-

tion du courant à véhiculer. En effet, le but étant de couper l'alimentation du circuit pendant certaines périodes, il est indispensable d'intégrer ce module. Les transistors utilisés dans ce cas sont des transistors communément appelés "ULL" pour Ultra Low Leakage. Ce sont des MOS extrêmement lents mais qui consomment extrêmement peu en comparaison avec les transistors Low Leakage ou High Speed. La consommation globale du circuit se résume donc aux uniques courants de fuite de ces transistors de coupure d'alimentation. Les courants de fuites de ce type de transistor sont de l'ordre de 30 à 40 fois inférieur à ceux des transistors "low leakage" et de 1300 à 1800 fois inférieur à ceux des transistors "high speed", en technologie 130nm. Ils sont de l'ordre de 5 à 6 fois inférieur à ceux des transistors "low leakage" et de 60 à 80 fois inférieur à ceux des transistors "high speed", en technologie 28 nm. Cet écart est dû au fait qu'il est très difficile de fabriquer des transistors qui consomment très peu dans les noeuds technologiques les plus avancés.

5.3.2.3 Processeur magnétique: implémentation

L'implémentation de ce processeur dans sa version CMOS/magnétique se décompose en plusieurs étapes:

- Code sources: à partir des codes sources du processeur CMOS, nous avons décrit un module "registre_mag" ayant un mot de 16 bits en entrée et un mot de 16 bits en sortie.
- Synthèse: la synthèse de ce type de circuit comportant à la fois des blocs CMOS et des blocs CMOS/magnétique doit se faire de façon hiérarchique et en 2 phases. Premièrement, il faut importer les sources du registre magnétique et en faire la synthèse logique mappée sur la bibliothèque comportant les bascules non volatiles. Deuxièmement, il faut importer l'ensemble des autres codes sources puis en faire la synthèse mappée sur les bibliothèques CMOS. A l'issue de ces 2 phases, le processeur non volatil est composée d'environ 200 flip-flops dont seulement 16 sont non volatiles.
- Placement et Routage: cette étape doit aussi se faire de façon hiérarchique. En effet, vu le faible nombre de bascules non volatiles, il est préférable mais pas indispensable, que toutes ces bascules soient alignées sur la même rangée. Cela permet entre autre de minimiser le nombre de ligne de champ d'écriture nécessaires et donc d'optimiser l'efficacité du routage de l'ensemble du circuit.

5.3.3 Etude de consommation selon plusieurs noeuds technologiques

Cette partie des travaux a pour objectif de montrer le gain potentiel que peut apporter une technologie hybride CMOS / magnétique. Pour cela, nous avons choisi l'application du processeur de démodulation numérique haute résolution présenté précédemment pour laquelle nous évaluons ce gain selon plusieurs noeuds technologiques. En ce qui concerne les procédés CMOS, nous avons étudié ce circuit dans les technologies STMicroelectronics 130nm, 65nm, 40nm et 28nm, dans leur version dites "high speed et "low power". En ce qui concerne les procédés magnétiques, nous avons étudié les technologies TAS et STT de finesses équivalentes aux finesses des procédés CMOS, en se basant sur des calculs théoriques. Cette étude est une projection vers l'avenir car aujourd'hui aucun fondeur n'est capable de fabriquer de façon industrielle des jonctions tunnel magnétiques de toute petite taille. Néanmoins, beaucoup d'efforts sont faits au niveau de la recherche dans ce domaine, notamment chez certaines sociétés industrielles, c'est pourquoi cette étude nous semble pertinente.

Dans notre application, le cycle est imposé par le convertisseur analogique / numérique double rampe à 16 bits. Ce cycle est de $10\mu s$. Le processeur effectue son calcul en maximum 70 cycles d'horloge. Nous avons choisi une horloge de 100 Mhz, soit 10ns de période. Le processeur effectue donc le calcul en 0.7 μs ce qui le laisse 9.3 μs inactif pour chaque échantillon, durée pendant laquelle le processeur peut être mis en état de veille et pour laquelle sont alimentation peut être coupée.

Pour faire cette étude, nous avons synthétisé les codes sources VHDL selon toutes les technologies STM mentionnées. Ceci nous a permis d'obtenir à partir des rapports de synthèse une estimation de la consommation statique. En effet, chacune des cellules de bibliothèque étant caractérisée précisémment par simulation électrique, et notamment les bibliothèques de STMicroelectronics [70], l'outil de synthèse est capable d'indiquer cette consommation en faisant la somme des courants de fuite de chaque cellule. Il est important de souligner que les modèles de simulation électrique servant à la caractérisation de cellules standards sont décrits selon le format standard BSIM3 pour le procédé 130nm, alors que pour les autres procédés que nous avons utilisés ils sont décrits selon le format standard BSIM4. Ce dernier prend en considération les aspects de courant de fuite de façon beaucoup plus précise et permet de modéliser aussi les courants de fuite dans les grilles, en plus de ceux dans les sources et drains [46]. Puis nous avons dans chaque cas extrait la "netlist" du circuit pour l'importer dans l'éditeur de schéma Cadence / Schematic Composer afin de faire une simulation électrique de l'ensemble du circuit. Un vecteur *lambda* a été appliqué sur chacune des entrées. Nous avons alors mesuré le courant statique consommé pendant

une phase de repos. Cette méthode, bien qu'à priori plus précise, s'est avérée peu fiable car le niveau de chaque noeud interne du circuit dépend fortement des stimulis d'entrée, qui sont eux-mêmes pas forcément très réalistes. Les résultats n'étant pas cohérents, ils ne sont pas présentés dans ce manuscrit.

Le premier tableau 5.2 de cette étude compare la consommation statique sur 1 cycle de conversion du CAN entre les technologies CMOS et le procédé CMOS / Magnétique TAS. Les constantes ne figurent pas dans ce tableau. Il s'agit du temps de chauffage des jonctions qui est de 10 ns, du nombre de registres à sauvegarder qui est 16, du champ de retournement qui est de 1 mA et du nombre de générateurs qui est de 1 pour 16 flip-flops.

Appl.	Process	JTM (nm)	I_{chauf} (μA)	E_{chauf} (pJ)	E_{chp} (pJ)	E_{MAG} (pJ)	E_{CMOS} (pJ)	$gain$ (%)
High Speed	130n (HS)	120	160	61	20	81	3208	97
	65n GP (SVT)	65	40	15	20	35	595	94
	40n LL (LVT)	40	15	5	20	25	15	Ø
	28n LL (LVT)	28	8	3	20	23	74	69
Low Leakage	130n (LL)	120	160	61	20	81	46	Ø
	65n LP (SVT)	65	40	15	20	35	2.4	Ø
	40n LS (SVT)	40	15	5	20	25	1.8	Ø
	28n LR (RVT)	28	8	3	20	23	5.6	Ø

TAB. 5.2 – *Etude de consommation statique d'un processeur simple CMOS vs CMOS/Magnétique TAS*

On peut remarquer à partir de ce tableau que l'énergie nécessaire à l'écriture des jonctions décroit lorsque les technologies sont de plus en plus avancées, mais faiblement. En effet, le courant nécessaire pour chauffer une jonction diminue sensiblement lorsque la taille de la JTM diminue alors que le courant nécessaire pour changer son aimantation ne diminue pas, car ce paramètre n'est pas fonction de la taille de la jonction.

On peut déduire de ce tableau qu'un procédé magnétique TAS permet de réduire de 97% et 94% la consommation statique sur des procédés CMOS 130n HS (High Speed) et 65n GP (General Purpose) SVT (Standard Vt), version dite "high speed" mais faible consommation. Effectivement le procédé 65n GP LVT (Low Vt) encore plus rapide ne figure pas dans cette étude. En ce qui concerne le procédé 40n LL (Low Leakage) LVT (Low Vt), un procédé hybride n'apporte rien car cette technologie est conçue pour des applications principalement "faible courant" mais de

rapidité moyenne. Il aurait été plus cohérent d'utiliser ici le procédé 40nm GP, mais nous n'avions pas à disposition ce PDK. Il en est de même pour le procédé 28n. En revanche, pour ce dernier, le procédé TAS permet tout de même de réduire la consommation statique de 69%, du fait que les courants de fuite augmentent fortement avec la miniaturisation des transistors.

En ce qui concerne les versions "low leakage" de ces procédés CMOS, très efficace en termes de réduction de la consommation, il est clair que le procédé TAS n'apporte pas d'intérêt du point de vue de la consommation statique. En revanche, être capable de déposer des jonctions assez petites, 65 nm ou 40 nm par exemple, sur une technologie mature donc moins onéreuse, 130 nm voire 180 nm, permettrait d'avoir un circuit non volatil, plus économe en énergie et à bas coût.

Le second tableau 5.3 de cette étude compare la consommation statique sur 1 cycle de conversion du CAN entre les technologies CMOS et le procédé CMOS / Magnétique STT. Les constantes ne figurent pas dans ce tableau. Il s'agit du temps d'écriture des jonctions qui est de 3 ns et du nombre de registres à sauvegarder qui est de 16 dans notre application.

Appl.	Process	JTM (nm)	I_{ecr} (μA)	t_{ecr} (ns)	E_{MAG} (pJ)	E_{CMOS} (pJ)	Gain (%)
High Speed	130n (HS)	120	226	3	13	3208	99
	65n GP (SVT)	65	40	3	3.8	595	99
	40n LL (LVT)	40	20	3	1.54	15	90
	28n LL (LVT)	28	12	3	0.58	74	99
Low Leakage	130n (LL)	120	226	3	13	46	72
	65n LP (SVT)	65	40	3	3.8	2.4	Ø
	40n LS (SVT)	40	20	3	1.54	1.8	14
	28n LR (RVT)	28	12	3	0.54	5.6	90

TAB. 5.3 – *Etude de consommation statique d'un processeur simple CMOS vs CMOS/Magnétique STT*

On peut déduire de ce tableau que l'énergie nécessaire à la sauvegarde des registres décroit fortement lorsque la taille de la jonction diminue, car le courant nécessaire diminue linéairement avec la taille de la JTM. Dans le même temps, le courant statique tend à augmenter dans les procédés CMOS submicroniques. On observe sur ce tableau que la consommation statique est légèrement plus faible pour le procédé 40nm que celle du procédé 65nm, ce qui n'est pas tout à fait cohérent avec la tendance. Nous avons alors analysé les architectures et nous nous sommes aperçu que

le nombre de cellules total diminue du procédé 65nm vers le procédé 40nm, et que la surface totale décroit fortement, -85%. En revanche, la surface diminue seulement de 60% pour le procédé 28nm par rapport à la génération précédente. Toutes ces différences peuvent s'expliquer entre autres par le fait que chacune des bibliothèques n'offre pas le même nombre de cellules standards complexes, ce qui a un impact direct sur la synthèse et les performances du circuit. Par conséquent, si l'on compare la consommation de ces 3 procédés par unité de surface, on retrouve bien une augmentation cohérente de la consommation statique depuis le procédé 65nm vers le procédé 28nm.

Ce tableau montre qu'une technologie hybride CMOS / magnétique STT permettrait de réduire de 90% à 99% la consommation statique du circuit processeur dans notre application de démodulation haute résolution, sur des procédés dit "high speed". En ce qui concerne les versions dites "low leakage", nous pouvons noter qu'un procédé magnétique STT apporte un gain en consommation de 72% en technologie CMOS 130nm, devient compétitif à partir de 40nm avec un gain de 14%, et permet de réduire de 90% la consommation statique en 28nm. Enfin, une technologie STT moyennement avancée, 65nm par exemple, permettrait de réduire la consommation statique de 32% d'un procédé CMOS 28nm, et encore plus sur des noeuds technologiques CMOS plus avancés.

5.4 Conclusion

L'évolution des procédés microélectroniques montre que la consommation statique devient un réel enjeu dans les systèmes d'aujourd'hui. C'est le cas notamment pour tous les appareils portables pour lesquels l'autonomie, entre autres, est une caractéristique essentielle. La tendance des procédés microélectroniques montre que les courants de fuite sont de plus en plus importants, surtout en dessous des noeuds technologiques 40nm, et qu'ils représentent une part de plus en plus importante dans la consommation totale d'un circuit intégré. Il est vrai qu'aujourd'hui les seuls produits industriels hybrides existant sont faits par la société Everspin sur les procédés magnétiques les plus anciens, FIMS, et que cela ne peut répondre ni aux besoins de faible consommation ni de miniaturisation. Cependant, cette même société annonce qu'elle sortira dans l'année 2012 une première mémoire industrielle sur la technologie STT. Ceci prouve que ce mode d'écriture représente un intérêt certain pour les prochaines années et cette étude montre que lorsque ces technologies seront disponibles à des tailles très petites alors nous pourrons envisager de réduire fortement la consommation des circuits intégrés. Par ailleurs, il n'est pas nécessaire d'avoir la

même "génération" de procédé entre le CMOS et le magnétique pour obtenir un gain significatif. Il est tout à fait possible de concevoir un circuit hybride avec un procédé CMOS 28nm et un post-process magnétique 65nm par exemple. Cette étude montre que cela permettrait tout de même de réduire la consommation statique d'environ 30%.

Par ailleurs, en dehors des aspects de la consommation, nous avons également montré que pour des applications pour lesquelles ni la consommation ni la vitesse ne sont des critères de performance, une technologie CMOS/magnétique a d'autres avantages. Etre capable de sauvegarder l'état d'un circuit à tout moment et pouvoir restaurer un état stable à n'importe quel moment en cas de coupure d'alimentation intempestive est un atout majeur. Ceci est possible grâce à la non volatilité apportée par l'utilisation de jonctions tunnel magnétiques dans les circuits CMOS.

Pour conclure ce chapitre nous pouvons dire que le fait d'avoir tous les outils et flots de conception disponibles pour concevoir un ASIC intégrant des jonctions tunnel ouvre des voies d'exploration. Cela permet de comparer en simulation le comportement d'un tel circuit avec son équivalent CMOS et d'évaluer le gain que cela peut apporter. Le gain en consommation peut être considérable dans des applications autonomes, comme dans la téléphonie mobile par exemple, où les appareils sont bien plus longtemps en mode de standby qu'en activité. Ces composants magnétiques amènent également aux réflexions sur les architectures de systèmes sur puces, car cette étude a été faite sans aucune considération système. Les résultats présentés dans ce chapitre illustrent surtout ce que l'on peut faire avec ces outils, en particulier en termes de consommation, ce qui n'est pas intégré aux simulateurs de très haut niveau. En effet, en plus d'avoir des registres non volatils, il est probable que certaines mémoires soient remplacées par des MRAMs, ce qui représente un gain en surface et en énergie. En effet, sauvegarder un contenu dans des mémoires distantes coûte en puissance et en temps. Un tel procédé CMOS/magnétique offre donc de nouvelles perspectives prometteuses.

Chapitre 6

Réalisation et tests de démonstrateurs

6.1 Introduction

Un des objectifs du projet de recherche SPIN était d'une part la mise en place d'une fabrication d'un procédé hybride et d'autre part la conception de démonstrateurs, intégrant de nouvelles architectures. Dans ce cadre, nous avons eu l'opportunité de concevoir et faire fabriquer plusieurs petits circuits de test, que nous présentons dans la suite de ce chapitre.

Par ailleurs, la société Crocus Technology travaille sur plusieurs applications de type mémoire. Ils fabriquent régulièrement des lots de plusieurs wafers afin d'améliorer et perfectionner leurs produits. Dans la majorité des cas, les améliorations sont faites au niveau du procédé de fabrication, et plus rarement au niveau du design. De fait, ils utilisent le même jeu de masques pour plusieurs fabrications successives. Nous avons appris en juin 2011 que Crocus travaillait sur de nouvelles architectures de leurs mémoires et qu'ils allaient utiliser un nouveau jeu de masques pour la prochaine fabrication. Crocus étant une startup du laboratoire CEA-Spintec, nous avons eu l'accord et l'opportunité d'intégrer sur leurs masques un démonstrateur. Nous avons donc accédé à leurs machines locales à Crocus-Grenoble, équipée du kit de conception de la technologie en question. Il s'agit du procédé CMOS 130n de TowerJazz et du procédé magnétique TAS de Crocus.

6.2 Démonstrateur SPIN: projet ANR

Dans le démonstrateur du projet SPIN, nous avons partagé une couronne de plots avec plusieurs partenaires, le LIRMM, l'IEF et Spintec. Le nombre de plots

dont nous disposions nous a permis d'intégrer 3 démonstrateurs, tous basés sur la bascule innovante que nous proposons, composée du latch magnétique non volatil ultra compact et d'un latch SRAM CMOS à 6 transistors classique. La figure 6.1 montre l'intégration de tous les démonstrateurs sur un seul circuit.

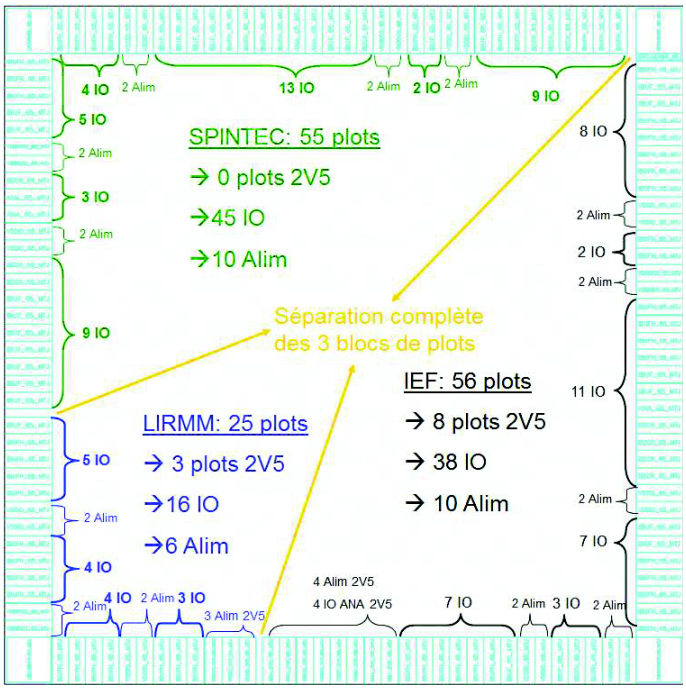


FIG. 6.1 – *Répartition de la couronne de plots pour le circuit commun du projet SPIN*

6.2.1 Registre à décalage non volatil

L'objectif de ce premier bloc est de valider sur silicium l'architecture que nous proposons, pour laquelle tout un ensemble de simulations électriques ont été faites. Ce registre à décalage est composé de 3 bascules. Chacune d'entre elles est commandée par les mêmes signaux magnétiques et la même horloge. Le schéma de ce bloc est présenté sur la figure 6.2.

Ce circuit a été simulé électriquement en utilisant le modèle compact TAS des jonctions tunnel. Le schéma utilisé pour ces simulations intègre également les plots d'entrée et de sortie utilisé pour le dessin des masques. Le comportement de ce circuit est montré sur la figure 6.3. Il a été vérifié du point de vue du LVS et DRC avant l'envoi en fabrication. On remarque sur ce résultat de simulation une alternance de fonctionnement CMOS classique notée "Décalage" avec une phase de restauration des données dans les jonctions, notée "Lecture". On peut noter que la valeur récupérée

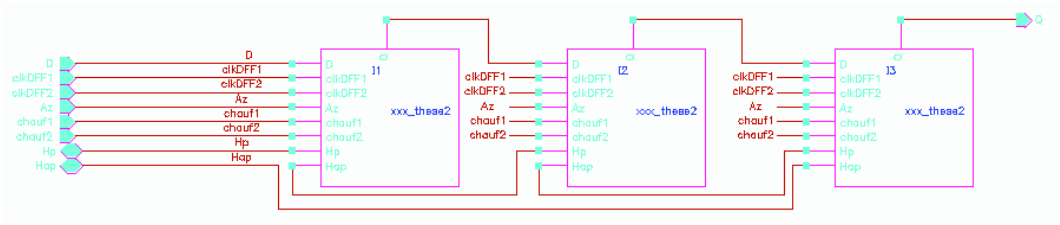


FIG. 6.2 – *Schéma du registre à décalage non volatil*

à chaque lecture est la même, soit "110". Suit une phase d'écriture pour laquelle le contenu de certaines jonctions est modifié, ce qui se vérifie à chaque lecture suivante ou la valeur est alors "010".

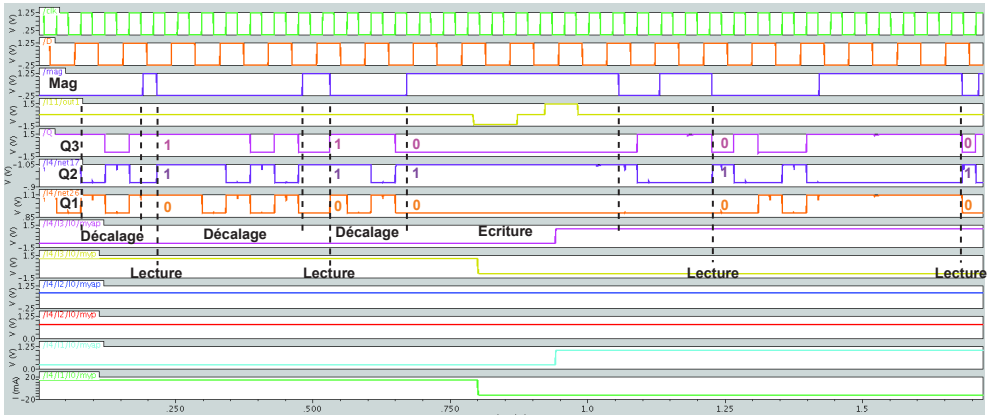


FIG. 6.3 – *Simulation du registre à décalage non volatil*

Le dessin des masques de ce circuit a été fait entièrement manuellement. Comme le montre la figure 6.4, toutes les cellules sont alignées les unes aux autres, de façon à ce que les rails d'alimentation soient en face les uns des autres, ainsi que la ligne de champ d'écriture.

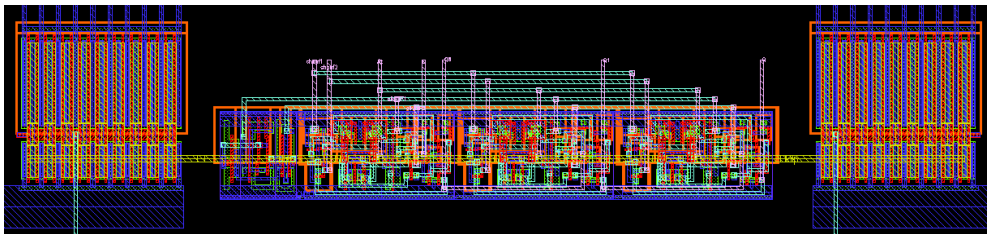


FIG. 6.4 – *Dessin des masques du registre à décalage*

Les générateurs de courant placés de part et d'autre du registre ont été dimen-

sionnés de façon à délivrer un courant de 20 mA. Cette valeur est typique, voire au-delà du besoin, et a été validée en simulation.

6.2.2 Compteur non volatil

L'objectif de ce second bloc est de démontrer le fonctionnement d'un circuit pour application haute sécurité, pour lequel les sorties des bascules sont sauvegardées à chaque cycle d'horloge. Ce circuit a premièrement été décrit en langage VHDL puis synthétisé en utilisant notre bascule non volatile. La "netlist" après synthèse a été importée sous Cadence / Schematic Composer, afin de simuler électriquement ce bloc, avec les plots d'entrée et de sortie. Ainsi, comme le montre la figure 6.5, on peut restaurer un état stable à n'importe quel moment en cas de coupure d'alimentation intempestive.

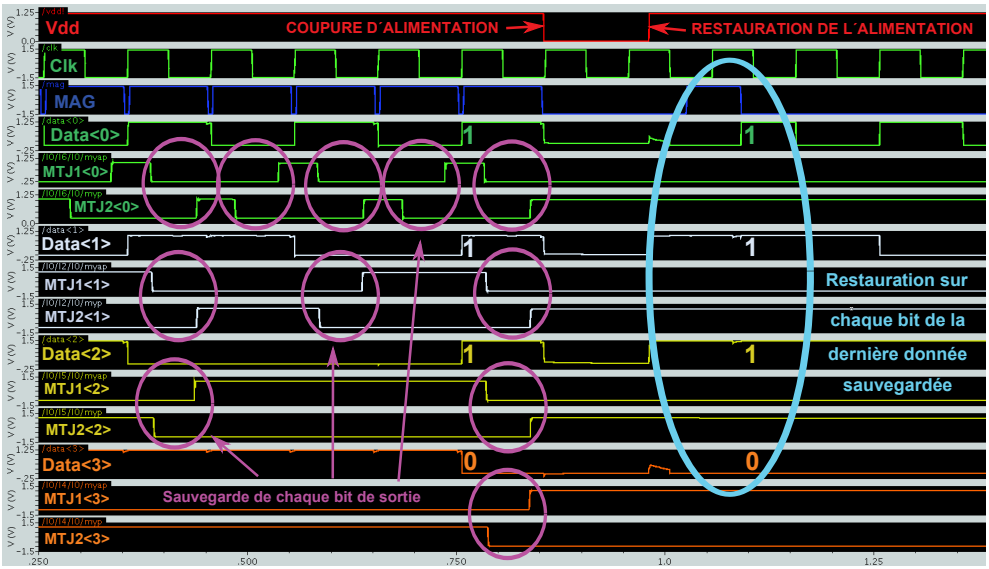


FIG. 6.5 – *Simulation du compteur non volatil*

Le dessin des masques a été fait sous Cadence / SOC Encounter. Les cellules ont été placées et routées automatiquement avec les outils présentés précédemment dans ce manuscrit. En revanche, à ce moment de la thèse, le flot de conception n'était pas finalisé donc les générateurs de courant ont été placés et routés manuellement, ce qui a été largement réalisable vu la complexité du circuit.

6.2.3 Machine à états non volatile

L'objectif de ce troisième bloc est de démontrer le fonctionnement d'un circuit non volatil, un peu plus complexe qu'un simple compteur composé de 5 cellules standard plus 4 bascules. Cette machine à états modélise le comportement de 2 feux tricolores. Elle possède 7 bascules non volatiles ainsi qu'une cinquantaine de cellules standards. De la même façon que pour le compteur, ce bloc a premièrement été décrit en langage VHDL, puis synthétisé en utilisant notre bascule non volatile. La "netlist" après synthèse a été importée sous Cadence / Schematic Composer, afin de simuler électriquement ce bloc, avec les plots d'entrée et de sortie, comme illustré sur la figure 6.6.

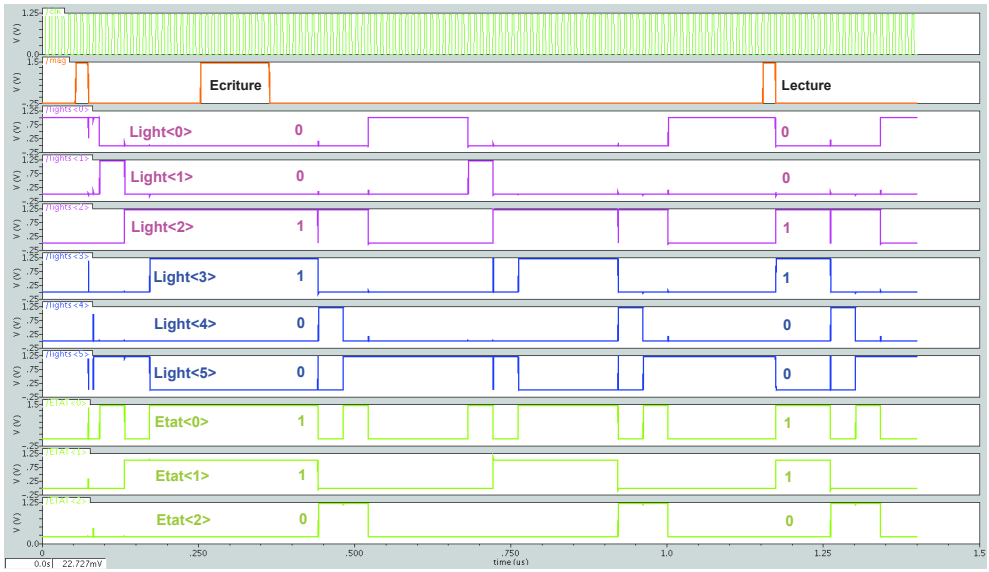


FIG. 6.6 – *Simulation de la machine à états non volatile*

Les signaux "light<0:2>" correspondent aux sorties des feux vert, orange et rouge du premier feu tricolore, les signaux "light<3:5>" correspondent aux sorties des signaux du second feu, et les signaux "Etat<0:2> représentent le codage de l'état de la FSM. Il s'agit de la phase de lecture. Comme le montre la figure 6.6, il est possible de revenir à n'importe quel moment dans le dernier état sauvegardé, aussi bien pour les signaux de sortie que pour les signaux de codage de l'état de la FSM.

Le dessin des masques a également été fait sous Cadence / SOC Encounter. Les cellules ont été placées et routées automatiquement et les générateurs placés et routés manuellement. La figure 6.7 montre la vue layout de ce bloc.

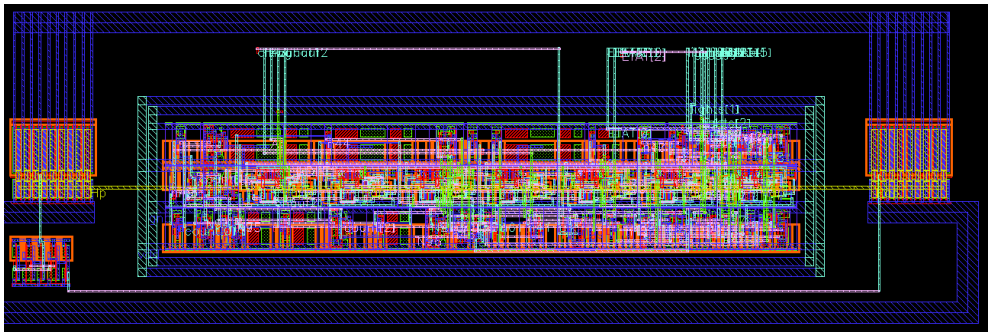


FIG. 6.7 – *Dessin des masques de la machine à état non volatile*

6.3 Test du démonstrateur SPIN

L'ensemble des tests fonctionnels que nous avons réalisés l'ont été sur un testeur industriel. En effet, Spintec a acquis en 2011 un testeur Diamond de chez LTX Credence. Il est destiné au test de circuits numériques, analogiques et signaux mixtes, avec la possibilité de tester des circuits prototypes pour le debug ou la caractérisation ainsi que des circuits produits en grand volume. Il peut contenir jusqu'à 10 cartes de test. Le testeur peut gérer jusqu'à 768 signaux numériques et 16 alimentations. Nous utilisons actuellement 2 cartes: la carte VIS16 pour les alimentations : celle-ci possède 16 sources de courant/tension quatre quadrants indépendantes. Les tensions max sont +/- 20V avec +/- 300mA. La carte DPIN96 permet de tester des circuits numériques. Jusqu'à 96 signaux numériques indépendants peuvent être testés. Pour chaque canal, les niveaux de tension, de courant, les timing, les formats et paramètres de mesure peuvent être contrôlés indépendamment. La tension max des signaux est 12V. Les tests fonctionnels sont programmés avec le langage Standard Test Interface Language (STIL). Par ailleurs, pour réaliser ces tests, nous avons conçu une carte de type PCB comportant un support de boitier, ainsi que des supports de type DIL autour. L'ensemble de cet environnement de test est illustré sur la figure 6.8. Le boîtier utilisé est un PGA 144. En effet, notre démonstrateur étant une partie intégrante du démonstrateur du projet SPIN, constitué de bloc d'autres partenaires, le nombre de signaux à câbler est proche de 144. Le plan de câblage de l'ensemble de ces blocs a été choisi arbitrairement et fourni à chacun des partenaires du projet ANR.

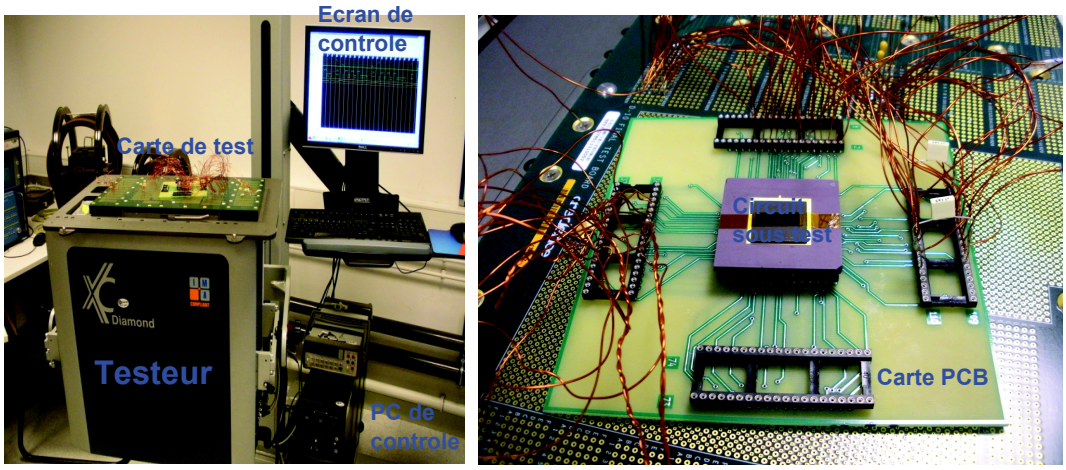


FIG. 6.8 – *Environnement de test*

6.3.1 Registre à décalage non volatil

Ce bloc ne comportant que 3 cellules flip-flop, il s'agit du plus simple à tester. Nous avons donc commencé nos tests par celui-ci. La première étape consiste à configurer le testeur en définissant chaque signal puis à les affecter à un canal du testeur. Les niveaux d'alimentation et de seuil de commutation Voh et Vol doivent également être définis ainsi que les chronogrammes des signaux appliqués sur les entrées. Il est donc nécessaire par la suite de faire une table de correspondance entre le brochage du PGA144 et de la carte d'interface du testeur afin de les relier physiquement avec des fils. Nous avons alors pu entreprendre les tests.

Nous avons commencé par tester la partie CMOS de ce registre, c'est à dire faire fonctionner le circuit sans phase ni de lecture ni d'écriture des jonctions tunnel. Au cours de ces premiers tests, aucun signal n'était modifié en sortie. Nous avons donc décidé de mettre en place un test de continuité sur tous les signaux, ce qui permet de s'assurer que la connexion physique est assurée. Le principe de ce test est d'envoyer un courant dans les plots et mesurer la tension. Si la connexion est ouverte alors il n'y a pas de tension, si elle est correcte alors on lit la tension de seuil des diodes de protection ESD intégrées aux plots. Ce test a permis d'identifier de nombreuses mauvaises connexions qui survenait de façon aléatoire. Pour régler ce problème, nous avons décidé de souder les fils sur la carte d'interface du testeur, prévue à cet effet. Nous n'avons alors plus rencontré ce type de problème.

Nous avons ensuite regardé les signaux d'entrée à l'oscilloscope. Nous nous sommes aperçu que les formes n'étaient pas très carrées et qu'il y avait un certain bruit. Nous avons alors décidé d'ajouter des capacités de découplage sur toutes les alimentations.

Notons d'ailleurs qu'il est indispensable d'alimenter tous les plots d'alimentation de la couronne de plot pour un bon fonctionnement, car en effet, si un ou plusieurs plots Vdd n'est pas alimenté, alors les diodes de protection ESD ne sont pas polarisées et deviennent passantes, ce qui dégrade complètement les signaux.

A l'issu de ces améliorations, nous avons constaté que le registre à décalage fonctionnait parfaitement sur plusieurs échantillons. Nous avons alors testé une phase de lecture des jonctions. Nous avons constaté qu'à une tension de 1.5V au lieu de 1.2V la lecture était stable et répétable, sans toutefois connaitre l'état des jonctions après fabrication. D'après les discussions que nous avons eues avec des personnes de Crocus, elles devraient toutes être dans le même état. La lecture n'est donc pas très cohérente dans la mesure où nous n'avons pas une jonction dans un état parallèle et l'autre dans un état antiparallèle.

Nous avons finalement introduit une phase d'écriture des jonctions avant une phase de lecture. Les tests ont montré que l'écriture ne se faisait pas. Plusieurs causes étaient possibles. Soit les jonctions ne sont pas assez chauffées et la température maximum est inférieure à la température de blocage, soit le champ généré n'est pas assez important. Le nombre de plots que nous avions à disposition ne nous permettait pas de séparer l'alimentation de la circuiterie et celle des générateurs. Nous avons donc augmenté la tension d'alimentation de l'ensemble du circuit dans le but de chauffer d'avantage les jonctions et de générer plus de champ. Les résultats de tests étaient identiques aux précédents. Sans pouvoir augmenter la tension d'alimentation indéfiniment nous souhaitions élever encore la température des jonctions, en chauffant l'ensemble du circuit. Pour cela nous avons utilisé une plaque chauffante régulée à température variable. Nous avons alors augmenté progressivement et par palier la température de cette plaque, jusqu'à 90°C environ. Malgré cela, nous n'avons pas réussi à lire la valeur attendue en sortie de nos bascules du registre. Nous avons alors essayé d'écrire les jonctions par un champ extérieur. Pour cela nous avons utilisé un aimant relativement puissant. Nous l'avons placé au-dessus du circuit, dans un sens pendant la première phase d'écriture puis dans le sens opposé pour la seconde phase. Malgré cela, il a été impossible de valider une lecture et une écriture selon les stimulis. Nous nous sommes alors retournés vers le LETI, en charge de certaines étapes du post-process et des tests préliminaires des jonctions. Il nous a été reporté que certaines étapes dégradaient le rendement de jonctions potentiellement fonctionnelles, mais que des mesures correctives avaient été prises pour le second lot de ce démonstrateur. Nous pouvons également préciser que pour l'instant, aucun partenaire n'a réussi à écrire des jonctions par la combinaison du chauffage et de la génération d'un champ d'écriture. La caractérisation qui est faite au LETI consiste à

changer l'aimantation de la couche de référence sous fort champ, quelques centaines d'Oersted, à température ambiante, dans le but d'extraire les valeurs de résistance et de TMR.

6.3.2 Compteur non volatil

Après avoir pris toutes les précautions identifiées pour le test de circuits intégrés, nous avons testé le compteur/décompteur non volatil. Au moment où nous avons fait ces tests, nous n'avions à disposition plus que 2 circuits car les autres ont été envoyés à un de nos partenaires, le LIRMM, pour effectuer les tests de leur partie. Parmi ces 2 circuits, l'un d'entre eux présentait un défaut de connexions. En effet, le test de continuité a révélé un circuit ouvert sur un signal d'entrée. Puisque ce phénomène n'apparaissait pas sur le second circuit, nous en avons déduit que le problème ne provenait pas des connexions entre la carte de test, le support et le testeur. Nous avons alors vérifié le câblage au microscope, mais aucune coupure sur les fils n'a pu être identifiée. Il s'agit certainement d'un mauvais contact métallique au niveau des plots de câblage ou un défaut de fabrication.

Nous avons alors testé le seul circuit qu'il nous restait, sachant que ni l'écriture ni la lecture n'avait fonctionné sur le démonstrateur du registre à décalage. De la même façon, le test de continuité a montré une mauvaise connexion sur un signal, mais de sortie cette fois-ci, ce qui nous a tout de même permis de faire le test du circuit. Dans un premier temps nous avons inhibé les signaux magnétiques pour tester la partie CMOS seulement. Nous avons constaté que le décompteur 4 bits fonctionnait parfaitement, sur les 3 bits observables. Lors de la mise sous tension toutes les sorties sont à 1 et à chaque front montant d'horloge, le mot de sortie décroit successivement. Nous avons alors dans un second temps activé les signaux pilotant la partie magnétique. Nous avons à nouveau constaté que la lecture est bien répétitive et que l'on récupère systématiquement la même donnée à la lecture. En revanche nous n'avons pas réussi à écrire dans les jonctions tunnel. De nouveaux tests seront faits sur le second lot de fabrication dans les prochaines semaines. Les wafers sont à ce jour au LETI pour les dernières étapes de fabrication.

6.3.3 Machine à états non volatile

Lors de ce test, nous avons constaté également 1 problème de continuité sur un signal d'entrée sans que nous ayons pu le résoudre, nous avons alors pu tester seulement 1 exemplaire. Les premiers résultats ont montré que toutes les sorties, soit 9 au total, oscillent à une fréquence de 1 GHz, quel que soit la fréquence de l'horloge,

typiquement 50 MHz. Il est donc impossible d'interpréter le niveau des signaux sur front de l'horloge. Une première hypothèse est que l'alimentation n'a pas le temps de s'établir de façon stable avant l'application des premiers stimulis, par effet induit des capacités de découplage. En effet, nous avons connecté 2 capacités en parallèle, une de 100 nF et une autre de 4,7 μF, permettant ainsi de découpler l'alimentation et de filtrer les fréquences hautes et basses. Nous avons déterminé par calcul qu'il faut environ 0.3 ms pour les charger avec un courant limité dans le circuit de 20 mA. Dans la mesure où notre environnement de test prévoit un temps de 10 ms pour la stabilisation des alimentations avant que le test ne commence, cette piste n'est pas la bonne pour résoudre ces phénomènes d'oscillation. La seconde piste consiste à connecter les capacités de découplage basses fréquences le plus près possible du circuit. L'idéal aurait été de prévoir une carte de test spécifiquement pour ce circuit et d'intégrer ces capacités de découplage sur le PCB. Comme cette carte de test est générique à plusieurs applications et que les alimentations ne sont pas situées aux mêmes endroits, cela n'est pas le cas. En revanche, nous avons pu souder ces capacités au niveau du support de test, soit très proche du circuit. En vain, car le résultat était strictement le même.

Pour améliorer ces phénomènes parasites, il est également conseillé de faire un plan de masse sur la carte de test, ce qui aurait éventuellement pu améliorer la situation si toutefois la raison était une perturbation extérieure. Dans la mesure où les circuits que nous testons ne fonctionnent pas à une fréquence très élevée, l'intérêt nous paraissait quasi nul. Par conséquent, nous n'avons pas réussi à faire fonctionner ce bloc lors des tests.

6.4 Démonstrateur Crocus: filière industrielle

Cette phase de conception s'est faite en un temps très limité car nous avons été informés de cette nouvelle fabrication à partir d'un nouveau jeu de masques relativement tardivement, peu de temps avant la date de départ en fabrication. La première étape était de prendre en main le kit de conception de la technologie CMOS et magnétique. La principale difficulté était de n'avoir que très peu de support de la part de Crocus, les personnes pouvant le faire étant aux Etats-Unis, certainement très occupées par la conception de leurs nouveaux circuits à implémenter dans ce même lot de fabrication. De plus, le nombre de licences pour les outils était limité. Les américains étant eux-mêmes aussi en phase de conception, avec des simulations très longues, il était régulier de ne pas avoir de licence disponible. Les conditions étaient donc difficiles et le temps alloué assez court.

Le choix des structures implémentées s'est fait en fonction du nombre de plots disponibles. Ces plots sont tout simplement un empilement de métaux et une ouverture de passivation, sans circuiterie autour, donc sans buffer ni d'entrée ni de sortie, et sans diode de protection. Leur taille est de $70\mu m \times 57\mu m$. Les personnes organisant le réticule final nous ont imposé d'avoir une seule rangée de 23 plots alignés, espacés de $47\mu m$. Les structures devaient être placées dans cette surface allouées. Nous avons donc décidé de supprimer 2 plots afin de placer 2 structures à leur emplacement, comme illustré sur la figure 6.9 représentant la vue layout finale de ce démonstrateur.

FIG. 6.9 – *Démonstrateur inclus dans le run Crocus Technology*

Ces 2 structures sont d'une part le latch magnétique non volatil que nous proposons, ainsi que la flip-flop associée. Le nombre de plots étant confortable, seul 3 signaux sont communs: Az, Chauf1 et Chauf2. Nous avons fait le choix de séparer les alimentations de chacun de ces blocs pour qu'ils soient tout à fait indépendant et que l'augmentation de l'un si besoin ne détériore pas l'autre. Nous avons également séparé les alimentations de la partie circuiterie de celle des générateurs. En effet, le CMOS fonctionne sous 1.2V alors que nous avons conçu des générateurs de courant en 3.3V. Une fois que nous avons pu établir le floorplan et distribuer les signaux en fonction des plots disponibles, nous avons pu concevoir chacune de ces structures.

6.4.1 Latch magnétique

Tout d'abord, rappelons que l'architecture du latch est composée à la fois de transistors ayant une tension de seuil haute pour les NMOS et des transistors ayant une tension de seuil basse pour les PMOS. La technologie CMOS de TowerJazz à laquelle nous avions accès ne permet pas d'utiliser des transistors avec différents Vt. En revanche, des transistors 1.2V et 3.3V étaient disponibles. Nous avons donc mixé le type des transistors afin d'avoir des PMOS qui fuient plus que les NMOS ce qui permet d'obtenir un "1" logique sur une des 2 branches du latch à l'état de repos, ou de maintien, ceci étant un des principe de base de notre montage. Les PMOS sont donc des transistors 1.2V et les NMOS sont des 3.3V. En effet, les MOS 3V3 ont des oxydes de grille plus épais, des Vt plus importants et fuient donc moins. L'association de ces types de transistors a été validée par simulations électriques.

Dans cette filière de fabrication, la taille des jonctions tunnel est fixe, de $200nm$.

Etant donné que plus la jonction a une taille importante plus le courant nécessaire pour la chauffer et important, et de façon quadratique, les tailles des transistors permettant de véhiculer ce courant sont loin d'être les tailles minimums de la technologie. En effet, la taille des PMOS est de $16\mu m/0.13\mu m$ et la taille des NMOS est de $26\mu m/0.13\mu m$. Nous avons alors simulé cette architecture électriquement, sans pour autant avoir la possibilité de faire des simulations Monte Carlo, le kit de conception ne le permettant pas. Nous avons alors fait des simulations paramétriques sur les tailles des transistors ainsi que sur la taille des jonctions. Après quelques itérations, nous avons déterminé une configuration robuste de ce latch magnétique. La phase suivante était alors l'implémentation du dessin des masques. Le layout de ce latch magnétique est présenté sur la figure 6.10.

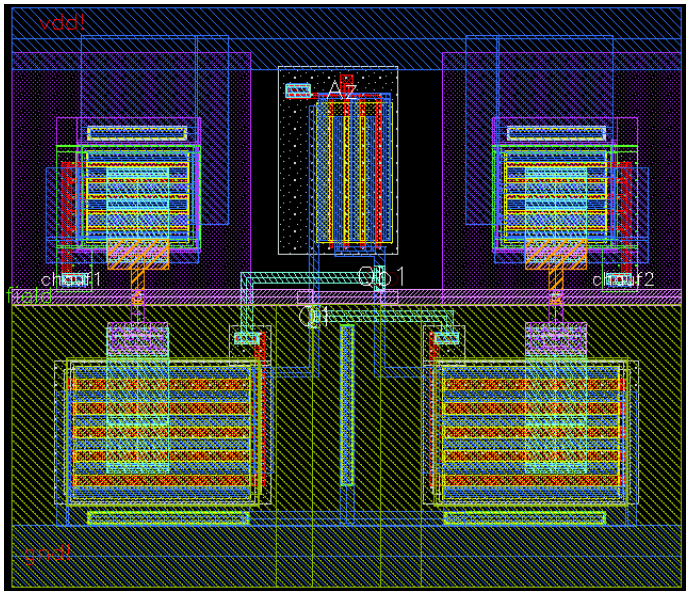


FIG. 6.10 – *Latch magnétique du démonstrateur inclus dans le run Crocus Technology*

6.4.2 Flip-Flop magnétique

Après avoir validé le latch magnétique, nous avons conçu la flip-flop, pour laquelle tous les autres transistors, d'accès et du latch CMOS, sont des transistors 1.2V. Les tailles des transistors d'accès sont les mêmes que celles du run SPIN ainsi que celle du latch CMOS qui a pu être dessiné avec des tailles de transistors proches des minimums. De la même façon, le comportement de cette flip-flop a été validé par simulation électrique en faisant des simulations paramétriques, ce qui a permis de passer à la dernière phase de conception, celle du dessin des masques. Le layout de cette bascule est présenté sur la figure 6.11.

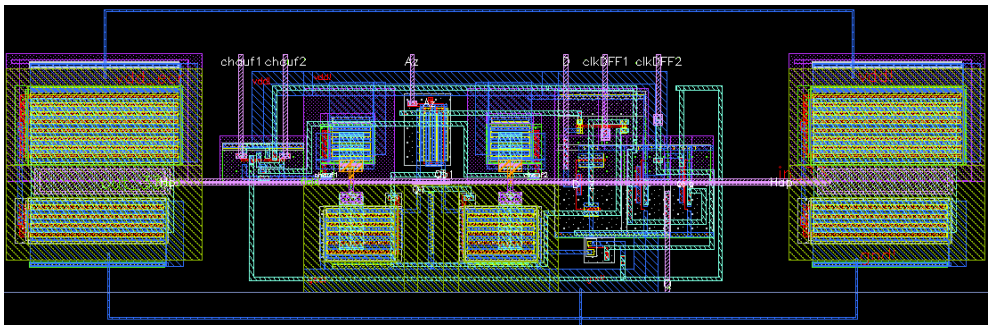


FIG. 6.11 – *Flip-Flop du démonstrateur inclus dans le run Crocus Technology*

6.5 Test du démonstrateur Crocus

Afin d'utiliser la même carte de test que pour le démonstrateur SPIN, nous avons choisi le même boitier, soit un PGA144. Ce boitier étant carré, il y a 36 plots par côté. Puisque notre structure ne comporte que 21 plots, nous avons placé 4 circuits dans la cavité du boitier pour optimiser le nombre de circuits testables par boîtier. Parmi ces 4 circuits, 2 sont ceux décrits dans ce manuscrit et 2 sont ceux d'un autre doctorant de Spintec qui a également conçu un démonstrateur dans ce run Crocus. Cette intégration et le plan de câblage sont illustrés sur la figure 6.12

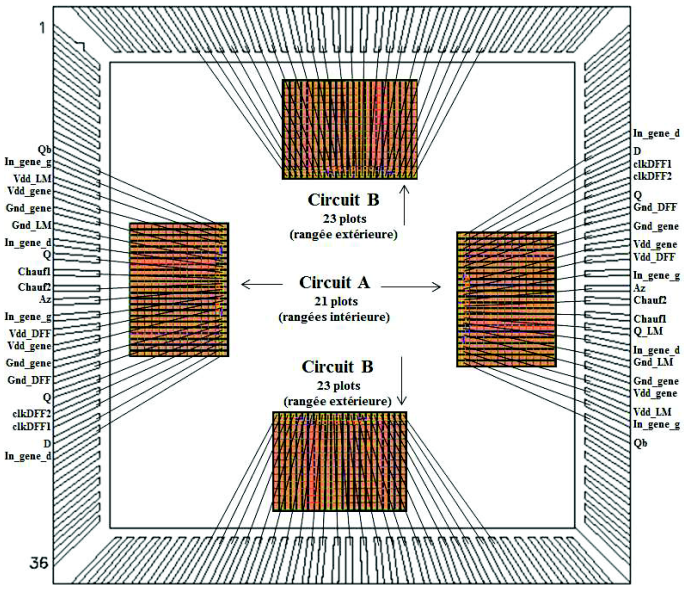


FIG. 6.12 – *Plan de câblage du démonstrateur inclus dans le run Crocus Technology*

Ce plan de câblage a été fait de façon à ce que les 2 circuits "A" puissent être testés avec les mêmes connexions sur le testeur, simplement en retournant le boîtier

de 180 degrés. Nous nous sommes alors assuré que la matrice de routage du PGA144 était parfaitement symétrique par rapport à l'origine géométrique.

6.5.1 Flip-Flop magnétique

L'expérience acquise lors des tests du circuit du run SPIN nous ayant servi, nous l'avons mise à profit pour le test de ce démonstrateur. Nous avons donc commencé les tests sur la flip-flop, composée d'entrées et de sorties faciles à piloter. L'objectif était dans un premier temps de valider le fonctionnement de la partie CMOS. Suite à l'application des premiers stimulis nous nous sommes aperçus que cette cellule ne fonctionnait pas. Nous avons vérifié quelques points sur le layout, tels que le nom des plots, leur positionnement, les connexions. Nous avons alors vu que nous avions mis sur tous les signaux des diodes de protection contre les antennes. Ceci nous a alors permis de faire un test de continuité et de nous apercevoir qu'un signal n'était pas transmis. Nous avons également vérifié la forme des signaux à l'oscilloscope, qui étaient conformes aux stimulis, bien qu'il y ait du bruit. Enfin, nous avons vérifié le câblage des circuits selon le plan de câblage que nous avions imposé, à l'aide d'un microscope. Ce câblage était conforme à notre demande. Malgré tout cela, nous n'avons pas obtenu de résultat concluant sur cette cellule. Nous avons donc décidé de tester le latch magnétique pour lequel nous avions accès à tous les signaux. C'était d'ailleurs ce à quoi nous avions pensé lors de la conception et du choix de l'implémentation des architectures.

6.5.2 Latch magnétique

De la même façon, nous avons axé les premiers tests sur la partie CMOS. Les 2 terminaux Q et Qb du latch peuvent être utilisés, de façon complémentaire, soit en entrée, soit en sortie. Nous avons donc imposé un signal carré sur Q et observé Qb, puis fait l'inverse. Dans les 2 cas, nous avons observé le même comportement, c'est à dire que d'imposer un '1' logique sur Q (ou Qb) permet bien d'écrire un '0' logique sur Qb (ou Q), en revanche imposer un '1' logique sur Q (ou Qb) ne permet pas d'écrire un '0' logique sur Qb (ou Q). Nous en avons déduit que le courant de fuite du PMOS n'est pas assez important pour avoir un niveau de tension proche de Vdd et que par conséquent le NMOS opposé n'est pas dans un état passant. Nous avons pu vérifier ceci en imposant un pulse sur le transistor PMOS supposé fuir suffisamment afin de le rendre passant un court instant, et effectivement, une fois le NMOS opposé passant, l'état reste stable. Grâce à cette astuce, nous avons pu avoir sur chacun des noeuds Q et Qb le niveau de tension attendu pendant toutes les phases de maintien,

de lecture et d'écriture. L'étape suivante était donc de tester la partie magnétique.

La première étape a été de configurer les stimulis selon les valeurs nominales déterminées en simulation. Quelle que soit la valeur écrite, la valeur lue est toujours la même, soit '0'. Les raisons potentielles que nous avons identifiées sont multiples. Soit le courant traversant les jonctions n'est pas assez important pour atteindre la température de blocage, soit le champ généré est trop faible. Nous avons alors commencé par augmenter la tension d'alimentation des générateurs, progressivement de 3.3V à 6V. Le résultat étant le même, nous avons alors augmenté progressivement la tension d'alimentation du latch, de 1.2V à 1.5V. Le comportement reste identique. Nous avons alors mis en place la mesure du courant dans les 2 alimentations, à partir d'un programme spécifique du testeur. Il s'est avéré que sur 6 circuits, pour 5 d'entre eux aucun courant ne circule dans le générateur de courant de champ d'écriture. Ceci est étrange et reste inexpliqué à ce jour. Il peut s'agir soit d'une erreur de programmation du testeur soit une rupture de la piste métallique si le courant est trop important. Pour le dernier circuit, nous avons bien observé une variation de courant de 15 mA à 45 mA en fonction de la tension d'alimentation. Nous avons alors observé le courant circulant dans le latch. La mesure a montré un courant de 1.3 mA pour une tension d'alimentation de 1.2V. Cette valeur est de l'ordre de grandeur du courant attendu par simulation puisque nous relevons un courant de 1.25 mA dans la branche du latch où la jonction est écrite, le reste pouvant être divers courant de fuite et parasites. Nous avons alors beaucoup travaillé sur les stimulis d'entrée afin de modifier certains paramètres essentiels, tel que le temps de chauffe des jonctions par exemple. Puis en faisant varier les seuils de détection des niveaux haut et bas, $V_{OH} et V_{OL}$, nous nous sommes aperçu que les délais de propagation étaient très longs. Nous avons alors cherché à comprendre les raisons. La première chose à laquelle nous avons pensé était les capacités parasites dues aux plots de câblage. En effet, cet empilement de tous les niveaux de métaux sur une telle surface représente une capacité d'environ 2 pF. Puisque nous n'avions plus vraiment accès aux machines de conception de Crocus, il était difficile de refaire de nouvelles simulations en intégrant ces paramètres. Nous les avons tout de même faites sur l'environnement de la fabrication pour le projet SPIN, ce qui nous a permis de nous rendre compte des nouveaux temps à définir sur les entrées. Puis en discutant avec un responsable d'équipe test chez STMicroelectronics, nous avons appris que les capacités parasites sur un tel testeur étaient de l'ordre de 10 à 15 pF. Nous avons alors établi de nouvelles simulations en intégrant ces nouveaux paramètres. Ces nouveaux stimulis ont été implémentés sur l'environnement de test. Cependant, il était toujours impossible d'écrire et/ou de lire une valeur différente. A ce moment-

là, nous n'étions pas certains de savoir si le problème était l'écriture ou la lecture. Nous avons alors eu l'idée de vérifier les niveaux de tension des noeuds Q et Qb lors d'une phase d'écriture. En effet, les deux résistances étant supposées dans un état opposé, leur valeur doit être différente. Les transistors étant les mêmes dans les deux branches du latch, la valeur de la résistance est le seul paramètre différent. Nous avons vérifié en simulation, que lors d'une phase d'écriture, le courant circulant dans une seule branche, le niveau de tension Q ou Qb de cette même branche est soit 0.18V soir 0.25V selon la valeur de la résistance. L'idée était donc de contrôler le niveau de tension de ces noeuds, en faisant varier le seuil de détection du niveau bas, avant et après une écriture. Lorsque nous avons fait ces mesures, la plupart des circuits que nous avions à disposition avaient déjà été bien éprouvés. Nous avons tout de même pu identifier, que le niveau de tension semblait différent sur ces noeuds Q et Qb après écriture de valeurs opposées successives. Néanmoins, nous ne sommes pas sûrs des précisions du testeur et le niveau de tension n'étant pas très stables, avec une tendance à décroitre, il n'est pas évident que notre analyse soit la bonne.

6.6 Conclusion

Que ce soit dans le cadre du projet SPIN ou du run de Crocus Technology, avoir la possibilité de concevoir plusieurs démonstrateurs a été une immense opportunité. D'une part, cela nous a permis d'implémenter nos structures innovantes dans des circuits plus ou moins simples et de les intégrer dans le flot de conception que nous avons mis en place. De plus, la conception en vue de la fabrication sensibilise sur plusieurs aspects auxquels on ne s'attarde pas en simulation. C'est le cas par exemple pour la structure de la couronne de plots, la séparation des alimentations, le choix des plots d'entrée et de sortie. C'est également le cas pour le dimensionnement d'une piste de métal dans lequel un fort courant circule, notamment pour les lignes de champ d'écriture.

Quelles conclusions et quels enseignements pouvons-nous tirer de la conception et le test de ces démonstrateurs ?

Concernant ceux du projet SPIN, le compteur ayant été synthétisé puis placé et routé selon le flot de conception que nous avons mis en place, nous pouvons dire que ce travail est validé, au moins sur une architecture simple de circuit intégré. De plus, bien que les résultats ne soient pas optimums, ces démonstrateurs ont permis tout de même de valider certains points. D'une part, l'un des deux brevets porte sur la phase de maintien du latch et de la bascule innovante sur une structure SRAM ultra compact à seulement 4 ou 5 transistors, sans résistance de charge, par différence de courant de fuite entre les transistors PMOS et NMOS. Dans la mesure où la partie CMOS et la lecture ont fonctionné pour plusieurs démonstrateurs, nous pouvons conclure que cette architecture et ce concept sont validés et conformes au comportement auquel nous nous attendions.

Concernant ceux du run Crocus, nous avons plusieurs interrogations car parallèlement aux tests que nous avons faits sur ce démonstrateur, les tests du circuit de l'autre doctorant de spintec ont également été faits. Il s'est avéré que plusieurs circuits ont fonctionné. Certes l'architecture de ce démonstrateur et le mode de fonctionnement sont différents, mais les procédés de fabrication CMOS et magnétiques sont les mêmes. Il est tout de même peu probable que toutes les jonctions des blocs que j'ai conçus ne soient pas fonctionnelles. Alors quelles peuvent être les raisons ? D'une part, l'architecture que j'ai implémentée fait appel à des notions de courant de fuite, ce qui est un élément clé du fonctionnement de ce dispositif. Il n'est pas du tout certain que les courants de fuites du procédé 130n soient très bien modélisés dans le kit de conception de TowerJazz. Les modèles sont certainement décrits selon

le standard BSIM3 comme pour le procédé 130n de STMicroelectronics. De plus, les jonctions ayant une taille très importante et le courant de chauffe étant quadratiquement proportionnel, les transistors ont dû être très gros, ce qui implique des capacités et comportement peut-être pas tout à fait ceux modélisés. Rappelons que le latch intègre des transistors 1,2V et des transistors 3,3V alors que l'ensemble du montage est alimenté en 1,2V. N'y a-t-il pas là une source de dysfonctionnement ? Par ailleurs, nous nous sommes aperçus que le test de continuité utilisé dans les diodes de protection ESD que nous avions dessinées n'est finalement pas très fiable dans la mesure où un circuit qui fonctionne montre des problèmes de continuité. Il n'est donc pas certain que le montage n'ait pas de problème de ce type. La difficulté de la phase de test par rapport à la phase de conception est d'une part le nombre d'échantillon disponible et d'autre part que nous n'avons pas accès aux paramètres de la température et de l'état magnétique.

Enfin, nous conclurons en disant que malgré les espoirs que nous avions, et malgré le nombre d'heures que nous avons passé au test de ces 5 démonstrateurs, cette étape a été très formatrice. D'une part d'un point de vue conception en vue du test mais d'autre part d'un point de vue test, ce qui sera je l'espère profitable pour le test de futurs circuits. Nous pouvons mentionner qu'à ce jour, le deuxième lot de fabrication SPIN est en fin de procédé de fabrication / découpe / assemblage, et que de nouveaux échantillons seront à tester dans les toutes prochaines semaines. Nous avons l'espoir que ceux-ci fonctionnent car d'après les informations que nous avons de la part du LETI, certaines étapes process critiques ont été améliorées.

Conclusion générale

Cette thèse s'est inscrite dans le cadre de deux projets de recherche ANR (Agence National de la Recherche), CILOMAG pour CIrcuits LOgiques MAGnétiques, de 2007 à 2010 et SPIN pour SPintronics for Innovative Nanotechnologies de 2009 à 2011. Nous avions plusieurs objectifs, tant au niveau conception, que du développement de kit de conception spécifique, que de l'étude comparative entre les technologies classiques de la microélectronique et technologies hybride CMOS / Magnétique.

Il était naturel de commencer par faire l'état de l'art des technologies non volatiles émergentes. Ce premier chapitre a permis de faire ressortir que le procédé MRAM - Magnetic Random Access Memory en utilisant la méthode d'écriture STT - Spin Transfer Torque était un axe de recherche à suivre et que les prochains développements d'applications se focaliseront particulièrement sur cette méthode d'écriture. Ceci a d'ailleurs été également soulevé par les experts de l'ITRS (International Technology Roadmap for Semiconductor's) des groupes ERD (Emerging Research Devices) et ERM (Emerging Research Materials), mentionnant que les technologies STT-MRAM et RedOx RAM étaient les plus prometteuses pour l'avenir.

Ensuite, il était indispensable de connaitre, comprendre et maitriser les flots de conception full custom et numérique, si différents l'un de l'autre, de même que les contraintes et les outils de conception CAO. Le second chapitre est donc exclusivement consacré à ces sujets. Il décrit chacune des étapes pour ces deux flots ainsi que leurs spécificités. Ce chapitre permet de situer la phase de conception dans le processus de fabrication, depuis la spécification jusqu'à son implémentation au niveau physique. Ce chapitre permet également de montrer les 2 grandes familles de circuits intégrés, analogique (ou full custom) et numérique.

Toujours dans un but de miniaturisation de circuits intégrés et des systèmes en général, nous avons étudié plusieurs architectures de cellules élémentaires magnétiques innovantes. L'état de l'art montre que les cellules existantes aujourd'hui sont basées sur le principe d'un point mémoire SRAM à 6 transistors et que l'intégration de jonctions tunnel selon le mode d'écriture TAS augmente le nombre de transistors de 3. Ce type de latch non volatil a donc au total 9 transistors. Nous proposons dans le troisième chapitre une nouvelle architecture basée sur le principe d'une cel-

lule mémoire à 4 transistors et sans résistance de charge, ce qui a amené à un dépôt de brevet d'invention. Une séquence d'écriture selon la méthode TAS très spécifique permet de ne pas augmenter le nombre de transistors malgré l'intégration de deux jonctions tunnel. Le nombre total de transistors est potentiellement de 4, voire 5 selon certaines technologies CMOS. Ce latch ultra compact a par ailleurs été intégré à un autre latch SRAM classique pour constituer une bascule non volatile, dans le but de l'intégrer au flot de conception numérique.

Dans un but de concevoir des circuits complexes en intégrant des composants magnétiques, nous avons développé un kit de conception complet. D'une part nous avons mis en place l'ensemble des éléments et des fichiers technologiques nécessaires pour concevoir des circuits selon le flot full custom, c'est à dire la simulation électrique, le dessin des masques simplifié grâce à une cellule paramétrable, l'extraction de tous les composants CMOS et magnétiques ainsi que les vérifications DRC et LVS. Cette première partie a été fournie à l'ensemble des partenaires des projets CILOMAG et SPIN et leur a permis d'élaborer de nouvelles architectures et de faire fabriquer des démonstrateurs respectant toutes les règles de conception et de fabrication. D'autre part, ont suivi le développement et la mise en place d'un flot de conception numérique, couvrant également l'ensemble des spécificités de ce flot : description comportementale, simulation électrique à partir d'un modèle Verilog, synthèse logique, placement / routage et vérification DRC et LVS. L'ensemble de ces travaux ayant été réalisés sur la technologie magnétique TAS, nous nous sommes également préoccupés du dimensionnement et de l'implémentation des générateurs de courant nécessaires à la génération du champ d'écriture des jonctions. Ce travail a donc permis d'ouvrir des perspectives en termes de conception de circuits complexes.

Le cinquième chapitre a été consacré à l'intégration de jonctions tunnel dans les circuits numériques pour montrer de quelle façon une telle technologie pouvait améliorer les performances ou les possibilités d'un circuit intégré. Nous avons tout d'abord conçu un circuit de filtrage numérique non volatil pour une application "haute sécurité", dans laquelle l'état du circuit est sauvegardé à chaque cycle d'horloge. Les simulations numériques ont montré qu'il était possible de restaurer un état stable à n'importe quel moment. Enfin, dans le but d'atteindre notre dernier objectif qui était d'évaluer et comparer les procédés CMOS actuels aux procédés hybrides CMOS / Magnétiques, nous avons conçu un processeur simple de cas d'étude, intégré à une chaine de démodulation de signal haute résolution. Cette étude a permis de montrer d'une part qu'une technologie TAS permet de réduire de façon non négligeable, de 69% à 97% dans notre application, l'énergie statique consommée par les composants CMOS, sur des technologies CMOS plutôt orienté vitesse, et d'autre part qu'une

technologie magnétique STT permet de réduire la consommation statique de notre circuit de 90% sur un procédé 28 nm dit faible consommation. Cette étude montre que l'avenir de la microélectronique peut rimer avec technologie magnétique.

Enfin le dernier chapitre illustre les tests et les résultats des démonstrateurs. En ce qui concerne le run SPIN nous avons vu que le premier lot a permis de valider un certain nombre de choses, à savoir une partie du flot de conception sur la synthèse et le placement / routage spécifique à un procédé hybride CMOS / magnétique. Il a également permis de valider un des brevets d'invention déposé, celui concernant la phase de maintien des niveaux de sortie par différence de courant de fuite en utilisant des transistors à seuil de commutation différents. Le second lot sera testé dans les prochaines semaines. Au vu des changements apportés à certaines phases de fabrication du procédé magnétique, nous avons espoir de pouvoir valider les phases de lecture et d'écriture de l'architecture que nous proposons. En ce qui concerne le run Crocus, il est vrai que certaines jonctions ont été écrites et lues sur un autre circuit que celui que j'ai conçu, mais nous pouvons tout de même noter des différences majeures entres cette architecture et celle que j'ai testée. Tout d'abord l'utilisation de transistors 1V2 et 3V3, puis le fait qu'il y ait de très grosses capacités parasites directement sur les noeuds que l'on observe, ainsi que le peu de fiabilité que nous avons sur la certitude que les connexions soient correctes. Il serait tout de même intéressant de reprendre ces tests avec du recul après avoir fait l'analyse que nous venons d'écrire ci-dessus, afin de mettre en place une éventuelle autre stratégie de test.

Perspectives

Les trois thèmes principaux de ces travaux de thèse ouvrent trois portes en vue des prochains mois et prochaines années. Le travail qui a été effectué pendant ces années de thèse a été principalement orienté vers la technologie CMOS / Magnétique des projets ANR, soit le procédé CMOS 130n de STMicroelectronics et le procédé magnétique TAS de Crocus Technology. Aujourd'hui, nous n'avons aucun procédé suffisamment mature ouvert à la fabrication pour pouvoir concevoir des circuits ou des démonstrateurs. Cependant, nous savons que l'avenir risque de s'orienter vers des **technologies magnétiques STT**, c'est pourquoi il est important de commencer à travailler sur de **nouvelles architectures de cellule standard**. En effet, pour ce premier axe de travail, la première partie de la conception est presque indépendante de la technologie. Seul le dimensionnement des transistors et la caractérisation en dépend. Il est tout à fait possible dans un premier temps de développer un ensemble de cellules élémentaires pour construire une bibliothèque "magnétique". Il serait également intéressant de réfléchir à **l'intégration de jonctions tunnel dans des cellules élémentaires combinatoires** afin de voir si cela peut également améliorer les performances d'un circuit intégré. C'est ce qu'a fait T. Hanyu de l'université de Tohoku en présentant une cellule "full adder" hybride [85]. Le second axe de travail pourrait être, lorsqu'un **procédé hybride CMOS / STT sera disponible** pour la fabrication, de **développer et mettre en place le kit de conception complet** correspondant, afin d'être capable de concevoir des circuits complexes selon des flots standards et avec des outils de conception industriels, de la même façon que nous avons pu le faire dans la dernière partie de cette thèse avec le procédé TAS. C'est certainement ce que l'on sera amené à faire au cours du projet ANR DIPMEM qui débutera à l'automne 2012, dans lequel le laboratoire CEA-Spintec et le service CMP seront partenaires, entre autres. Ceci ouvrirait de nouvelles perspectives pour le troisième axe de travail, qui serait de **concevoir un circuit complexe CMOS / STT**, dans une application donnée, et de comparer les performances soit en termes de sécurité soit en termes de gain énergétique, selon l'application choisie, avec un ASIC CMOS. Un partenariat avec des spécialistes de la conception de circuit intégré dédié à la gestion de l'énergie, qui maitrisent les techniques de "power gating" par exemple, pourrait être une collaboration très intéressante. De la même façon, il serait très intéressant de travailler sur un **système complet**, afin d'évaluer le gain de performances grâce à l'intégration de jonctions

tunnel sur de nouvelles architectures où les mémoires pourraient soit être remplacées par des MRAM, soit supprimés au profit de bascules non volatiles au niveau cellules standards. C'est d'ailleurs un des axes de recherche du projet MARS auquel participe le laboratoire CEA-Spintec. D'un point de vue personnel, je serai impliqué, de près ou de loin, dans tous les projets mentionnés ci-dessus, notamment les projets DIPMEM et MARS. En effet, je serai affecté dès le 1er octobre 2012 au laboratoire Spintec du CEA qui participe notamment à un projet européen SPOT basé sur l'effet Spin-Orbit, dans lequel je serai très impliqué, sur de nouveaux composants magnétiques à 3 pôles, permettant d'une part de décorréler l'écriture et la lecture, et d'autre part de réduire fortement l'énergie nécessaire au retournement de la couche de stockage magnétique puisque l'écriture se fait sans qu'aucun courant ne traverse la barrière tunnel. Par conséquent, les prochaines années offrent de très intéressantes perspectives pour le domaine de la spintronique.

Brevets et Publications

Brevets

VOLATILE/NON-VOLATILE MEMORY CELL, Yoann Guillemenet, Lionel Torres, Gregory Di Pendina, Guillaume Prenat, Kholdoun Torki

LOADLESS VOLATILE/NON-VOLATILE MEMORY CELL, Gregory Di Pendina, Guillaume Prenat, Kholdoun Torki

Publications

Ultra Compact Non-volatile Flip-Flop for Low Power Digital Circuits based on Hybrid CMOS/Magnetic Technology, In Proceedings of Power And Timing Modeling, Optimisation and Simulation - PATMOS, Madrid, Spain, September 2011.

A Hybrid Magnetic/CMOS Process Design Kit for the Design of Low-power Non-volatile Logic Circuits, 56th Magnetism and Magnetic Materials Conference, Scottsdale, Arizona, USA, 30 October - 3 November 2011, in Journal of Applied Physics.

Hybrid CMOS/Magnetic Process Design Kit and application to the design of high-performances non-volatile logic circuits, ICCAD - International Conference on Computer-Aided Design, San Jose, California, USA, November 2011, Invited paper.

Hybrid CMOS/Magnetic Process Design Kit and application to the design of reliable and low-power non-volatile logic circuits, MEDIAN 2012, Annecy, France, June 2012.

Magnetic Process Design Kit for Hybrid CMOS / Magnetic process, University Booth, DATE - Design And Test in Europe conférence, Grenoble, France, April 2011.

Références

[1] The flash memory can be reprogrammed up to 100 times. *H8S/2357 Group - H8S/2357F-ZTATTM - H8S/2398F-ZTATTM Hardware Manual - Section 19.6.1. Renesas. 2004-10. Retrieved 2012-01-23.*, 2004.

[2] Samsung introduces the next generation of nonvolatile memory - pram. September 11th 2006.

[3] http://www.synopsys.com.cn/information/snug/2007-2008-collection/synopsys-power-gating-design-methodology-based-on-smic-90nm-process, 2007-2008.

[4] Ivar giæver - nobel prize winner in physics 1973, May, 19th 2010.

[5] Memristor faq. *Hewlett-Packard*, September 3rd 2010.

[6] Advanced solid-state memory systems and products: Emerging non-volatile memory technologies, industry trends and market analysis. Innovative Research and Products Inc., April 2011.

[7] IBM develops instantaneous memory 100x faster than flash. *engadget*, June 30th 2011.

[8] Projet ANR (Agence Nationale de la Recherche), Contrat ANR-06-NANO-066.

[9] Projet ANR (Agence Nationale de la Recherche), Contrat ANRSPIN (NANOINNOV-RT - 2009).

[10] K. Agarwal, K. Nowka, H. Deogun, and D. Sylvester. Power gating with multiple sleep modes. In *Proceedings of the 7th International Symposium on Quality Electronic Design*, ISQED '06, pages 633–637, Washington, DC, USA, 2006. IEEE Computer Society.

[11] AMD. Amd dl160 and dl320 series flash: New densities, new features. *AMD. 2003-17. Retrieved 2012-01-23. "The devices offer single-power-supply operation (2.7 V to 3.6 V), sector architecture, Embedded Algorithms, high performance, and a 1,000,000 program/erase cycle endurance guarantee."*, October 2003.

[12] M. Baibich, J. Broto, A. Fert, F. N. V. Dau, F. Petroff, P. Etienne, G. Creuzet, A. Friederich, and J. Chazelas. *Phys. Rev. Lett. 61*, (2472), 1988.

[13] T. Barnes. Skill: a cad system extension language. *Design Automation Conference*, pages 266–271, 1990.

[14] B.Dieny, A.Deac, M.Kerekes, O.Redon, J. Nozieres, U.Ebels, L.Prejbeanu, and R.Sousa. New write schemes for magnetic non-volatile memories: thermally assisted and spin transfer writing. *Trends in Nanotechnology*, 2006.

[15] L. Benini, A. Bogliolo, and G. D. Micheli. A survey of design techniques for system-level dynamic power management. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems - Special section on low-power electronics and design*, 8, N° 3, June 2000.

[16] L. Berger. Low-field magnetoresistance and domain drag in ferromagnets. *Journal of Applied Physics*, 49:2156–2161, 1978.

[17] L. Berger. Domain drag effect in the presence of variable magnetic field or variable transport current. *Journal of Applied Physics*, 50:2137–2139, 1979.

[18] L. Berger. Emission of spin waves by a magnetic multilayer traversed by a current. *Physical Review B*, 54:9353–9358, 1996.

[19] K. Bernstein, C.-T. Chuang, R. Joshi, and R. Puri. Design and cad challenges in sub-90nm cmos technologies. In *Computer Aided Design, 2003. ICCAD-2003. International Conference on*, pages 129 – 136, nov. 2003.

[20] G. Binash, P. Grünberg, F. Saurenbach, and W. Zinn. *Phys. Rev. B 39*, (4828), 1989.

[21] W. C. Black and B. Das. Programmable logic using giant-magnetoresistance and spin-dependent tunneling devices. *Journal of Applied Physics*, 87(9)(6674), 2000.

[22] S. Borkar. Design challenges of technology scaling. *Micro, IEEE*, 19, N°14:23–29, July-August 1999.

[23] D. A. Buck. Ferroelectrics for digital information storage and switching. *Report R-212, MIT*, June 1952.

[24] B. P. C. Marmé Thompson. *Emerging Memories. Technologies and Trends.* Number 0306475537. Kluwer Academic Publishers, 2002.

[25] B. Calhoun and A. Chandrakasan. Ultra-dynamic voltage scaling using sub-threshold operation and local voltage dithering in 90nm cmos. *IEEE International Solid-State Circuits Conference. Digest of Technical Papers. ISSCC*, 1,10:300–599, Feb 10th 2005.

[26] C. Chappert, A. Fert, and F. V. Dau. The emergence of spin electronics in data storage. *Nature Mater*, 11:813–823, November 6th 2007.

[27] W. Chedid and C. Yu. Survey on power management techniques for energy efficient computer systems.

[28] W. Cheng and B. Baas. Dynamic voltage and frequency scaling circuits with two supply voltages. *Circuits and Systems, ISCAS. IEEE International Symposium*, pages 1236–1239, May 2008.

[29] O. L. Chua. Memristor-the missing circuit element. *IEEE Transactions on Circuit Theory*, CT-18, 1971.

[30] P. Dandumont. La flash nor produite en flux tendu. *tom's Hardware, source EETimes*, May 10th 2010.

[31] R. Dennard, 1968.

[32] B. Dieny, V. S. Speriosu, S. S. P. Parkin, B. A. Gurney, D. R. Wilhoit, and D. Mauri. Giant magnetoresistive in soft ferromagnetic multilayers. *Phys. Rev. B*, 43:1297–1300, Jan 1991.

[33] L. Engelbrecht, A. Jander, and P. Dhagat. A toggle mram bit modeled in verilog-a. In *Semiconductor Device Research Symposium, 2009. ISDRS '09. International*, pages 1 –2, dec. 2009.

[34] G. Estrin. Organization of computer systems-the fixed plus variable structure computer. In *Western Joint Computer*, pages 33–40, New York, 1960.

[35] G. Estrin. Reconfigurable computer origins: the ucla fixed-plus-variable (f+v) structure computer. *Annals of the History of Computing, IEEE*, 24(4):3–9, Oct. - Dec. 2002.

[36] A. Fert and I. A. Campbell. *Phys. Rev. Lett. 21*, 1968.

[37] P. P. Freitas and L. Berger. Observation of s-d exchange force between domain walls and electric current in very thin permalloy films. *Journal of Applied Physics*, 57:1266–1269, 1985.

[38] D. Frohman-Bentchkowsky. Digest of technical papers. *ISSCC*, page 80, 1971.

[39] fr.wikipedia.org/wiki/Loi_de_Moore.

[40] M. Gerardin. Compte rendue de l'académie des sciences. 53:727, 1861.

[41] I. Giaever. Energy gap in superconductors measured by electron tunneling. *Physical Review Letters*, 5:147–148, August 1960.

[42] J. Grollier, V. Cros, A. Hamzic, J. M. Georgea, H. Jaffrès, A. Fert1, G. Faini, J. B. Youssef, , and H. Legall. Spin-polarized current induced switching in co/cu/co pillars. *Applied Physics Letters*, 78:3663, 2001.

[43] Y. Guillemenet, G. D. Pendina, G. Prenat, L. Torres, and K. Torki. Volatile/non-volatile memory cell, January 2011.

[44] J. A. Halderman, S. D. Schoen, N. Heninger, W. Clarkson, W. Paul, J. A. Calandrino, A. J. Feldman, J. Appelbaum, and E. W. Felten. Lest we remember: Cold boot attacks on encryption keys. *Center for Information Technology Policy, Full research paper Appeared in Proc. 17th USENIX Security Symposium (Sec 08), San Jose, CA, July 2008.*, 2008.

[45] M. Hosomi, H. Yamagishi, T. Yamamoto, K. Bessho, Y. Higo, K. Yamane, H. Yamada, M. Shoji, H. Hachino, C. Fukumoto, H. Nagao, and H. Kano. A novel nonvolatile memory with spin torque transfer magnetization switching: spin-ram. *International Electron Devices Meeting, IEDM Technical Digest. IEEE International*, pages 459–462, December 5th 2005.

[46] http://cmosedu.com/cmos1/BSIM4_manual.pdf.

[47] http://en.wikipedia.org/wiki/Cadence_SKILL.

[48] http://en.wikipedia.org/wiki/EPROM.

[49] http://en.wikipedia.org/wiki/Phase change_memory.

[50] http://en.wikipedia.org/wiki/System_on_a_chip.

[51] http://fr.wikipedia.org/wiki/%C3%89lectromigration.

[52] http://fr.wikipedia.org/wiki/Circuit_logique_programmable.

[53] http://fr.wikipedia.org/wiki/Radio identification.

[54] http://leom.ec lyon.fr/dea/DEA_DEI/Mod%E9lisation%20des%20transistors%20MOS.pdf.

[55] http://nvmw.ucsd.edu/2010/documents/Driskill Smith_Alexander.pdf : Grandis.

[56] http://onlyzentv.blogspot.com/2011/01/le-monde-de-la-spintronique electrons.html.

[57] http://tima.imag.fr/publications/files/rr/adc_234.pdf.

[58] http://www device.eecs.berkeley.edu/bsim/?page=BSIM3.

[59] http://www device.eecs.berkeley.edu/bsim/?page=BSIM4.

[60] http://www device.eecs.berkeley.edu/bsim/?page=BSIMCMG.

[61] http://www device.eecs.berkeley.edu/bsim/?page=BSIMSOI.

[62] http://www.electroniques.biz/mobile/article_interview.php/?id_article=406722.

[63] http://www.elec.ucl.ac.be/enseignement/ELEC2550/submicron MOS.pdf.

[64] http://www.esiee.fr/ francaio/enseignement/version_pdf/VII_CAN.pdf.

[65] http://www.everspin.com/CES2012/CES_2012_Everspin_Slide_Show.pdf.

[66] http://www.everspin.com/overview.php.

[67] http://www.fujitsu.com/cn/fsp/services/memory/fram.

[68] http://www.futura-sciences.com/fr/definition/t/informatique 3/d/pram_5198/.

[69] http://www.itrs.net/Links/2010ITRS/2010Update/ToPost/ERD_ERM_2010FINALReportMemoryAssessment_ITRS.pdf.

[70] http://www.mentor.com/products/ic-manufacturing/news/eldo simulator.

[71] http://www.neel.cnrs.fr/UserFiles/file/physique_pour_tous/article_de_revue/RefletsALbertFERT15_5 10.pdf.

[72] http://www.nims.go.jp/apfim/halfmetal.html.

[73] http://www.sigen.net/semi_soi.html.

[74] http://www.tech faq.com/eeprom.html.

[75] S. Ikeda, J. Hayakawa, Y. Ashizawa, Y. Lee, K. Miura, H. Hasegawa, M. Tsunoda, F. Matsukura, and H. Ohno. Tunnel magnetoresistance of 604diffusion in cofeb/mgo/cofeb pseudo-spin-valves annealed at high temperature. *Applied Physics Letters 93*, 8(082508), 2008.

[76] I.L.Prejbeanu, W. Kula, K. Ounadjela, R. Sousa, O. Redon, B. Dieny, and J. Nozieres. Thermally assisted switching in exchange-biased storage layer magnetic tunnel junctions. *IEEE Transactions on Magnetics*, 40(4), 2004.

[77] R. C. Johnson. 'missing link' memristor created: Rewrite the textbooks? *EETimes*, 2008.

[78] M. Julliere. *Physics Letters 54A, 225*, 1975.

[79] B. Loegel and F. Gautier. *J. Phys. Chem. Sol. 32*, (2723), 1971.

[80] Y.-H. Lu and G. D. Micheli. Comparing system-level power management policies. *IEEE Design & Test of Computers*, 18, N°2, March 2001.

[81] J. J. Makwana and D. K. Schroder. A nonvolatile memory overview.

[82] M. Marcoe. History of memory ram. *eHow*.

[83] F. Masuoka, M. Assano, H. Iwahashi, T. Komuro, and S. Tanaka. A new flash eeprom cell using triple polysilicon technology. *IEDM Tech. Dig.*, 30:464–467, 1984.

[84] S. Matsunaga, J. Hayakawa, S. Ikeda, K. Miura, T. Endoh, H. Ohno, and T. Hanyu. Mtj-based nonvolatile logic-in-memory circuit, future prospects and issues. *Design, Automation and Test in Europe Conference*, (ISBN: 978-3-9810801-5-5), 2009.

[85] S. Matsunaga, J. Hayakawa, S. Ikeda, K. Miura, T. Endoh, H. Ohno, and T. Hanyu. Mtj-based nonvolatile logic-in-memory circuit, future prospects and

issues. In *Design, Automation Test in Europe Conference Exhibition, 2009. DATE '09.*, pages 433 –435, april 2009.

[86] S. Matsunaga, J. Hayakawa, S. Ikeda, K. Miura, H. Hasegawa, T. Endoh, H. Ohno, and T. Hanyu. Fabrication of a nonvolatile full adder based on logic-in-memory architecture using magnetic tunnel junctions. *Applied Physics Express 1*, (091301), August 2008.

[87] L. Mearian. Is nand flash memory a dying technology? *Techworld. Retrieved 2010-02-04.*, February 4th 2010.

[88] M.ElBaraji, V. Javerliac, W.Guo, G.Prenat, and B. Dieny. Dynamic compact model of thermally assisted switching magnetic tunnel junctions. *Journal of Applied Physics*, 106(12):123906, 2009.

[89] Y. Monnet, M. Renaudin, R. Leveugle, S. Dumont, and F. Bouesse. An asynchronous des crypto-processor secured against fault attacks. *International Conference on Very Large Scale Integration (VLSI-SOC)*, pages 21–26, 2005.

[90] J. S. Moodera, L. R. Kinder, T. M. Wong, , and R. Meservey. Large magnetoresistance at room temperature in ferromagnetic thin film tunnel junctions. *Physical Review Letters*, 74(16):3273–3276, 1995.

[91] G. E. Moore. Cramming more components onto integrated circuits. *Electronics*, 38, April 1965.

[92] N. Mott. volume A 153, 1936.

[93] E. B. Myers, D. C. Ralph, J. A. Katine, R. N. Louie, and R. A. Buhrman2. Current-induced switching of domains in magnetic multilayer devices. *Science*, 285 no. 5429:867–870, August 6th 1999.

[94] K. Noda, K. Matsui, K. Takeda, and N. Nakamura. A loadless cmos four transistor sram cell in a 0.18 um logic technology. *Electron Devices, IEEE Transactions on*, 48(12):2851–2855, December 2001.

[95] S. S. P. Parkin and al. Giant tunnelling magnetoresistance at room temperature with mgo (100) tunnel barriers. *Nat. Mat. 3*, 12:862–867, 2004.

[96] A. Pavlov and M. Sachdev, editors. *CMOS SRAM. Circuit Design and Parametric Test in Nano-Scaled Technologies*. Springer, 2008.

[97] G. D. Pendina, G. Prenat, and K. Torki. Loadless volatile/non-volatile memory cell, January 2011.

[98] D. Ralph and M. Stiles. Spin transfer torques. *Journal of Magnetism and Magnetic Materials*, 320:1190–1216, April 2008.

[99] J. Reed and M. Bellis. Inventors of the modern computer: The invention of the intel 1103 - the world's first available dram chip. *Inventors.about.com*.

[100] O. Rossetto, P. Seen, and G. Sicard. *Conception analogique: cours de Master CSINA*. 2004.

[101] G. Rostky. Remembering the prom knights of intel. *EETimes*, 2002.

[102] R. Sandeep, N. Deshpande, and A. Aswatha. Design and analysis of a new loadless 4t sram cell in deep submicron cmos technologies. *2nd International Conference on Emerging Trends in Engineering and Technology (ICETET)*, pages 155–161, Dec 2009.

[103] L. Savtchenko, B.Engel, N. Rizzo, M. Deherrera, and J. Janesky. Method of writing to scalable magnetoresistance random access memory element, 8th 2003.

[104] L. Scheick, S. Guertin, and G. G.M. Swift. Analysis of radiation effects on individual dram cells. *Nuclear Science, IEEE Transactions on*, 47 (6):2534–2538, 2000.

[105] B. Sheu, D. Scharfetter, P. Ko, and M. Jeng. Bsim: Berkeley short-channel igfet model for mos transistors. *IEEE Journal of Solid-State Circuits*, 22 , n°4:558–566, August 1987.

[106] C. Sie, A. Pohm, P. Uttecht, A. Kao, and R. Agrawal. Chalcogenide glass bistable resistivity memory. *IEEE*, MAG-6:592, September 1970.

[107] C. Sie, R. Uttecht, H. Stevenson, J. D. Griener, and K. Raghavan. Electric-field induced filament formation in as-te-ge semiconductor. *Journal of Non-Crystalline Solids*, 2:358–370, 1970.

[108] C. H. Sie. Memory devices using bistable resistivity in amorphous as-te-ge films. *PhD dissertation, Iowa State University, Proquest/UMI publication 69-20670*, January 1969.

[109] J. Slonczewski. Current-driven excitation of magnetic multilayers. *Journal of Magnetism and Magnetic Materials*, 159:L1–L7, June 1996.

[110] E. C. Stoner and E. P. Wohlfarth. A mechanism of magnetic hysteresis in heterogeneous alloys. *Philosophical Transactions of the Royal Society A: Physical, Mathematical and Engineering Sciences*, 240(826):599–642, May 4th 1948.

[111] B. D. Strukov, S. G. Snider, R. D. Stewart, and R. S. Williams. The missing memristor found. *Nature*, 453:80–83, 2008.

[112] F. Tabrizi. The future of scalable stt-ram as a universal embedded memory. *Grandis Inc. - EETimes*, february 21th 2007.

[113] K. Takeda, Y. Aimoto, N. Nakamura, H. Toyoshima, T. Iwasaki, K. Noda, K. Matsui, S. Itoh, S. Masuoka, T. Horiuchi, A. Nakagawa, K. Shimogawa,

and H. Takahashi. A 16 mb 400 mhz loadless cmos four transistor sram macro. *Journal of Solid-State Circuits, IEEE*, 35(11):1631–1640, November 2000.

[114] A. Tal. Nand vs. nor flash technology: The designer should weigh the options when using flash memory. *Electronic products*, Retrieved 2010-07-31., February 2002.

[115] P. M. Tedrow and R. Meservey. Spin-dependant tunneling into ferromagnetic nickel. *Physical Review Letters*, 26(4):192–195, January 1971.

[116] M. Tsoi, A. G. M. Jansen, J. Bass, W.-C. Chiang, M. Seck, V. Tsoi, and P. Wyder. Excitation of a magnetic multilayer by an electric current. *Physical Review Letters*, 80:4281–4284, May 11th 1998.

[117] J. Wang and B. H. Calhoun. Canary replica feedback for near-drv standby vdd scaling in a 90nm sram. *Custom Intergrated Circuits Conference (CICC)*, 2007.

[118] W.Guo, G.Prenat, V. Javerliac, M.ElBaraji, N. DeMestier, C. Baraduc, and B. Dieny. Spice modeling of magnetic tunnel junctions written by spin-transfer torque. *Journal of Physics D: Applied Physics*, 43(21):215001, 2010.

[119] R. S. Williams. How we found the missing memristor. *IEEE Spectrum*, January 2008.

[120] D. C. Worledge. Spin flop switching for magnetic random access. *Applied Physics Letters*, 84-22:4559, 2004.

[121] D. C. Worledge. Single-domain model for toggle mram. *IBM Journal of Research and Development*, 50-1:69–79, January 2006.

[122] D. Wouters. Oxide resistive ram: challenges and potential for scaled memory application. *International Nanotechnology Conference INC*, May 2010.

[123] www.elec.ucl.ac.be/enseignement/ELEC2550/submicron MOS.pdf.

[124] W.Zhao, E. Belhaire, V. Javerliac, C.Chappert, and B. Dieny. A non-volatile flip-flop in magnetic fpga chip. In *IEEE International Design and Test of Integrated Systems conference*, Tunisia, 2006.

[125] J. Yang and L. Chen. A new loadless 4 transistor sram cell with a 0.18 um cmos technology. *Canadian Conference on Electrical and Computer Engineering Electrical CCECE*, pages 538–541, April 2007.

[126] Y.Guillemenet, L. Torres, G.Sassatelli, and N. Bruchon. On the use of magnetic rams in field-programmable gate arrays. *Int. Journ. of Reconfigurable Computing*, 2008.

[127] B. Yu, X. Sun, S. Ju, D. Janes, and M. Meyyappan. Chalcogenide-nanowire-based phase change memory. *Nanotechnology, IEEE Transactions*, 7(4):496–502, July 2008.

[128] G. Q. Zhang and A. Roosmalen, editors. *More than Moore Creating High Value Micro/Nanoelectronics Systems*. Springer, 2009.

[129] W. Zhao, E. Belhaire, and C. Chappert. Spin-mtj based non-volatile flip-flop. In *IEEE International Conference on Nanotechnology (IEEE-NANO)*, number ISBN: 978-1-4244-0607-4, pages 399–402, Hongkong, China, 2007.

[130] W. Zhao, E. Belhaire, V. Javerliac, C. Chappert, and B. Dieny. A non-volatile flip-flop in magnetic fpga chip. *Design and Test of Integrated Systems in Nanoscale Technology International Conference*, pages 323–326, September 2006.